Digitized by the Internet Archive
in 2025

https://archive.org/details/mind_2017_126_501-502

GRADUATE THEOLOGICAL UNION
LIBRARY
1962

MIND

A Quarterly Review of Philosophy

Joint Editors: A. W. Moore
Lucy O'Brien

Reviews Editor: Anil Gomes

Associate Editors: William Child
Luke Fenton-Glynn
Fiona Leigh
Stephen Mulhall
A. C. Paseau
Daniel Rothschild
Amia Srinivasan

Editorial Board: Nomy Arpaly, David Chalmers, David Christensen, Ursula Coope, Steven Crowell, Cian Dorr, James Dreier, Dorothy Edgington, Branden Fitelson, Moira Gatens, Warren Goldfarb, Sally Haslanger, Pamela Hieronymi, Richard Holton, Axel Honneth, Frank Jackson, James Joyce, Jennifer Lackey, Christian List, Ofra Magidor, Sarah Moss, Jennifer Nagel, Gideon Rosen, Sherri Roush, Jennifer Saul, Geoffrey Sayre-McCord, T. M. Scanlon, Daniel W. Smith, Julia Tao, Kenneth Taylor, Amie L. Thomasson, Timothy Williamson, Jonathan Wolff

Managing Editor: Stephanie Sacharov

Vol. 126 · No. 501 · January 2017

Published for the Mind Association by
Oxford University Press

Articles

Ramsey without Ethical Neutrality: A New Representation Theorem
EDWARD ELLIOTT 1

Hume's Doxastic Involuntarism
HSUEH QU 53

Cognitive Mobile Homes
DANIEL GRECO 93

Paternalism and our Rational Powers
MICHAEL CHOLBI 123

Metanormative Contextualism and Normative Uncertainty
JOHN PITTARD AND ALEX WORSNIP 155

Truth is Simple
LEON HORSTEN AND GRAHAM E. LEIGH 195

Discussions

Moral Judgement, Agency and Affect: A Response to Gerrans and Kennett
CHRIS ZARPENTINE 233

Mental Time Travel, Dynamic Evaluation, and Moral Agency
PHILIP GERRANS AND JEANETTE KENNETT 259

Book Reviews

Graham Priest: *One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*
MICHAEL PRICE 269

Wolfgang Spohn: *The Laws of Belief: Ranking Theory & its Philosophical Applications*
A. C. PASEAU 273

Henry Allison: *Kant's Transcendental Deduction*
SACHA GOLOB 278

Dominic Scott: *Levels of Argument: A Comparative Study of Plato's* Republic *and Aristotle's* Nicomachean Ethics
KAREN MARGRETHE NIELSEN 289

Nissim Francez: *Proof-Theoretic Semantics*
IVO PEZLAR 299

Alva Nöe: *Strange Tools: Art and Human Nature*
JOHN HYMAN 304

Ernest Sosa: *Judgment and Agency*
MIRIAM SCHLEIFER MCCORMICK 309

Erratum 319

Please visit the journal's websites at: http://www.york.ac.uk/depts/phil/mind/ and https://academic.oup.com/mind

Vol. 126 · No. 501 · January 2017

© Mind Association 2017. All rights reserved; no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without prior written permission of the Publishers, or a licence permitting restricted copying issued in the UK by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1P 9HE.

Environmental and ethical policies

Oxford Journals, a division of Oxford University Press, is committed to working with the global community to bring the highest quality research to the widest possible audience. Oxford Journals will protect the environment by implementing environmentally friendly policies and practices wherever possible. Please see http://www.oxfordjournals.org/ethicalpolicies.html for further information on environmental and ethical policies.

Digital object identifiers

For information on dois and to resolve them, please visit www.doi.org.

Permissions

For information on how to request permissions to reproduce articles/information from this journal, please visit www.oupjournals.org/permissions.

Advertising

Inquiries about advertising should be sent to Linda Hann, 1 Fernham Road, Shellingford, Oxfordshire, SN7 7PU. Email: lindaj.hann@btinternet.com. Tel: +44 (0)1367 710022.

Disclaimer

Statements of fact and opinion in the articles in Mind are those of the respective authors and contributors and not of *MIND* or Oxford University Press. Neither Oxford University Press nor *MIND* makes any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

Articles

Russellian Acquaintance and Frege's Puzzle
DONOVAN WISHON 321

Maximally Consistent Sets of Instances of Naive Comprehension
LUCA INCURVATI AND JULIEN MURZI 371

What Verities May Be
IGOR DOUVEN AND LIEVEN DECOCK 385

Deflationism, Arithmetic, and the Argument from Conservativeness
DANIEL WAXMAN 429

Able to Do the Impossible
JACK SPENCER 465

Everything, and Then Some
STEPHAN KRÄMER 499

Naïve Realism In Kantian Phrase
ANIL GOMES 529

Discussions

Preservation, Commutativity and *Modus Ponens*: Two Recent Triviality Results
JAKE CHANDLER 579

Supporters and Underminers: Reply to Chandler
RICHARD BRADLEY 603

Preservation and Postulation: Lessons from the New Debate on the Ramsey Test
HANS ROTT 609

Book Reviews

Eric Funkhouser: *The Logical Structure of Kinds*
JOSEPH LAPORTE 627

Gail Fine: *The Possibility of Inquiry: Meno's Paradox from Socrates to Sextus*
DAVID BRONSTEIN 631

Bence Nanay: *Aesthetics as Philosophy of Perception*
OPHELIA DEROY 635

Nils Holtug: *Persons, Interests, and Justice*
TIM CAMPBELL 643

Martin Smith: *Between Probability and Certainty: What Justifies Belief*
KELLY BECKER 647

Please visit the journal's websites at: http://www.york.ac.uk/depts/phil/mind/ and https://academic.oup.com/mind

Vol. 126 · No. 502 · April 2017

© Mind Association 2017. All rights reserved; no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without prior written permission of the Publishers, or a licence permitting restricted copying issued in the UK by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1P 9HE.

Environmental and ethical policies
Oxford Journals, a division of Oxford University Press, is committed to working with the global community to bring the highest quality research to the widest possible audience. Oxford Journals will protect the environment by implementing environmentally friendly policies and practices wherever possible. Please see http://www.oxfordjournals.org/ethicalpolicies.html for further information on environmental and ethical policies.

Digital object identifiers
For information on dois and to resolve them, please visit www.doi.org.

Permissions
For information on how to request permissions to reproduce articles/information from this journal, please visit www.oupjournals.org/permissions.

Advertising
Inquiries about advertising should be sent to Linda Hann, 1 Fernham Road, Shellingford, Oxfordshire, SN7 7PU. Email: lindaj.hann@btinternet.com. Tel: +44 (0)1367 710022.

Disclaimer
Statements of fact and opinion in the articles in Mind are those of the respective authors and contributors and not of *MIND* or Oxford University Press. Neither Oxford University Press nor *MIND* makes any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

Ramsey without Ethical Neutrality: A New Representation Theorem

EDWARD ELLIOTT
Australian National University
Edward.elliott@anu.edu.au

Frank Ramsey's 'Truth and Probability' sketches a proposal for the empirical measurement of credences, along with a corresponding set of axioms for a (somewhat incomplete) representation theorem intended to characterize the preference conditions under which this measurement process is applicable. There are several features of Ramsey's formal system which make it attractive and worth developing. However, in specifying his measurement process and his axioms, Ramsey introduces the notion of an ethically neutral proposition, the assumed existence of which plays a key role throughout Ramsey's system. A number of later representation theorems have also appealed to ethically neutral propositions. The notion of ethical neutrality has often been called into question—in fact, there seem to be good reasons to suppose that no ethically neutral propositions exist. In this paper, I present several new, Ramsey-inspired representation theorems that avoid any appeal to ethical neutrality. These theorems preserve the benefits of Ramsey's system, without paying the cost of ethical neutrality.

1. Introduction

In his posthumously published 'Truth and Probability', Frank Ramsey (1931b) sketches a proposal for the empirical measurement of credences, along with a corresponding set of axioms for a (somewhat incomplete) representation theorem intended to characterize the preference conditions under which this measurement process is applicable. Ramsey's formal approach is distinctive, first deriving a utility function to represent an agent's utilities, and then using this to construct their credence function. In specifying his measurement process and his axioms, Ramsey introduces the notion of an *ethically neutral proposition*, the assumed existence of which plays a key role throughout his system.

The existence of such propositions has often been called into question. Ramsey's own definition of ethical neutrality presupposes the philosophically suspect theory of logical atomism. On other common ways of defining the notion, it is frequently noted that we lack good

Mind, Vol. 126 . 501 . January 2017
doi:10.1093/mind/fzv180 Advance Access publication 14 October 2016

© Elliott 2016

reasons for supposing that ethically neutral propositions exist, and in some cases we find that there are very good reasons for supposing that they cannot exist. Any system which relies on the existence of such propositions ought to be rejected.

However, we can have a representation theorem which is essentially Ramseyian in character that does not succumb to these objections. In §4, I will develop several new representation theorems inspired by Ramsey's own which avoid appeal to ethically neutral propositions in any problematic sense, substantially improving upon Ramsey's proposal. In §5, I will note several features that the new theorems share with Ramsey's system which make them independently attractive, especially in comparison to later works such as Savage (1954) and Jeffrey (1983). First, however, I will discuss Ramsey's original proposal in some detail (§2), which will help us see why Ramsey thought he needed to introduce the notion of ethical neutrality, why it is problematic, and how those problems can be circumvented (§3).

2. Ramsey's proposal

2.1 Background

One of Ramsey's main goals in 'Truth and Probability' was to argue that the laws of probability supply for us the 'logic of partial belief' (1931b, p. 166). His argument proceeds by first attempting to say what credences *are*, and on the basis of that understanding, showing that they conform to the laws of probability.

Regarding the first step, Ramsey clearly had operationalist sympathies, asserting that the notion 'has no precise meaning unless we specify more exactly how it is to be measured' (1931b, p. 167). To be measured as having probabilistically coherent credences *is* (more or less) on this picture to have probabilistically coherent credences, and anyone who can be measured through Ramsey's procedure at all will have credences conforming to the laws of probability.[1] Note that the

[1] Setting operationalism aside, it is easy to see in 'Truth and Probability' an early statement of something like functionalism: degrees of belief are to be understood through their causal role with respect to behaviour when considered in conjunction with a total desire state. Ramsey writes that 'the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it' (1931b, p. 169). Ramsey argues against characterizing degrees of belief in terms of some introspectively accessible feeling had by a subject upon considering the relevant proposition. These arguments go well beyond operationalism, though I will not recapitulate them here. He concludes that 'intensities of belief-feelings ... are no doubt interesting, but ... their practical interest is entirely due to their

 © Elliott 2016

procedure was intended to be applicable to ordinary agents — Ramsey was not trying to define degrees of belief for some ideally rational being, but for the everyday person on the street (albeit not without some unavoidable idealization).

With this goal in mind, Ramsey proposes to take as the theoretical basis of his measurement system a particular theory of decision making — that is, the theory that 'we act in the way that we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions' (1931b, p. 173). His idea is to assume the basic truth of something like subjective expected utility theory, and on that assumption, use empirical information about an agent's preferences to work out what her credences and utilities must be. Ramsey was entirely aware of the empirical difficulties that this theory faces, writing that:

> [it] is now universally discarded, but nevertheless comes, I think, fairly close to the truth in the sort of cases with which we are most concerned ... This theory cannot be made adequate to all the facts, but it seems to me a useful approximation to the truth particularly in the case of our self-conscious or professional life, and it is presupposed in a great deal of our thought. (1931b, p. 173)

We will return shortly to exactly what Ramsey meant by 'the sort of cases with which we are most concerned', and exactly what he needed to assume to get his measurement process off the ground.

At several points Ramsey notes that a similar kind of reasoning is used frequently by physicists and other hard scientists in the development of systems for the measurement of non-psychological quantities. Measurement systems are never developed in a theoretical vacuum; the actual data we receive are always interpreted through the lens of some presupposed theory or another — and quite frequently, as Ramsey notes, such a theory might 'like Newtonian mechanics ... still be profitably used even though it is known to be false' (1931b, p. 173), so long as it is accurate for the cases at hand. Indeed, at several points Ramsey notes that measurement 'cannot be accomplished without introducing a certain amount of hypothesis or fiction ... if it is allowable in physics it is allowable in psychology also' (1931b, p. 168; cf. Krantz *et al* 1971, pp. 26–31 on the role of idealizing assumptions in the construction of measurement systems). For Ramsey, this hypothesis or

---

position as the hypothetical causes of beliefs *qua* bases of action' (1931b, p. 172). On this more charitable interpretation, Ramsey's representation theorem can be seen as spelling out precisely the relevant causal roles associated with credence states.

 © Elliott 2016

fiction is that ordinary folk are more or less expected utility maximizers, at least in the right circumstances.

With this in mind, we can summarize Ramsey's measurement procedure as follows:

(a) determine an agent's preferences over outcomes and gambles;

(b) define a relation of equal difference in utilities;

(c) locate ethically neutral propositions of probability ½;

(d) construct an interval scale representation of the agent's preferences;

(e) use that representation to define a probability function, which is taken to represent the agent's credences.

I will discuss each step in turn. For the sake of simplicity, I have neglected to discuss one important aspect of Ramsey's procedure — *viz.*, the use of preferences over complex gambles to define conditional probabilities, which are used to show that the measured credences constitute a probability function. (This part of Ramsey's procedure is outlined in Bradley 2001.) It is also worth emphasizing that much of what follows is a rational reconstruction — Ramsey's own remarks are sketchy at best, and he rarely explains the motivations for any of the steps he makes.

2.2 Determining a preference ordering

The first stage of Ramsey's procedure is to determine the agent's preferences. We begin with the agent's preferences over what I will call *outcomes*. This is, according to Ramsey, relatively straightforward:

> If ... we had the power of the Almighty, and could persuade our subject of our power, we could, by offering him options, discover how he placed in order of merit all possible courses of the world. In this way all possible worlds would be put in an order of value ... (1931b, p. 176)

Ramsey here makes it clear that he takes outcomes to be 'different possible totalities of events ... the ultimate organic unities' (1931b, pp. 177–8); that is, possible worlds. It is unclear here exactly what Ramsey intended by 'possible worlds'. For instance, it is unclear whether we are considering metaphysically or conceptually possible worlds, and whether we are only interested in worlds consistent with history up to the event of the choice.

 © Elliott 2016

In any case, we will designate the set of outcomes with $\boldsymbol{W} = \{w_1, w_2, \ldots\}$. Importantly, within only a few paragraphs of referring to the outcomes as worlds, Ramsey goes on to note that with respect to at least one proposition P, and some outcomes w_1 and w_2, '$[w_1]$ and $[w_2]$ must be supposed so far undefined as to be compatible with both P and $\neg P$' (1931b, p. 178, Fn. 1). The most natural interpretation of this seems to be that w_1 and w_2 ought to be considered *not quite* as worlds, but rather as propositions maximally specific with respect to everything except P. (See §3.1 for further discussion.) Generalizing, we can take Ramsey's outcome space $\boldsymbol{W}$ to be a set of very highly specific propositions, some — but not all — of which may be maximally specific. We will say that $w_1 \succ w_2$ just in case the agent prefers w_1 to w_2; $w_1 \sim w_2$ just in case the agent is indifferent between the two options. The first task in Ramsey's measurement procedure is, then, to determine a preference ordering over $\boldsymbol{W}$.

We are required also to empirically determine how the agent ranks *gambles*. Once again, Ramsey asks us to imagine that we have convinced our subject of our power, but this time we make offers of the following kind: 'Would you rather have world $[w_3]$ in any event, or world $[w_1]$ if P is true, and world $[w_2]$ if P is false?' (1931b, p. 177). Let us represent the latter option, the gamble *w_1 if P is true, and w_2 if P is false*, as simply $(w_1, P; w_2)$.[2] With his background assumption in mind, Ramsey then notes that:

> If… [the agent] were certain that P was true, he would simply compare $[w_1]$ and $[w_3]$ and choose between them as if no conditions were attached; but if he were doubtful his choice would not be decided so simply. (1931b, p. 177)

Here, Ramsey looks to compare an outcome with a gamble, so we are to assume that gambles and outcomes are comparable. It is also evident from the axioms he later provides that we need to consider agents' preferences between gambles. In sum, if $\boldsymbol{G}$ is the set of all gambles, then Ramsey requires us to empirically determine a $\succcurlyeq$-ordering on $\boldsymbol{W} \cup \boldsymbol{G}$.

[2] It is unclear exactly how we ought to understand Ramsey's gambles. For reasons outlined in Joyce (1999, pp. 62–3), 'w_1 if P is true, and world w_2 if P is false' should not be understood using material conditionals. It is not obvious whether either subjunctive or indicative conditionals would fare any better; though Sobel (1998, p. 239) suggests that $(w_1, P; w_2)$ is just a conjunction of subjunctives, $(P \mathbin{\square\!\!\rightarrow} w_1) \,\&\, (\neg P \mathbin{\square\!\!\rightarrow} w_2)$. Part of the exegetical difficulty here is due to Ramsey's lack of specificity regarding the nature of the outcome set, and to the interpretive difficulties present in Ramsey's proposed system discussed shortly below.

 © Elliott 2016

There are a number of interpretive difficulties with Ramsey's proposed system that might be raised at this point. For one thing, it is unclear how outcomes as highly specific as Ramsey suggests can be 'offered' to any ordinary human subject; the power to conceptualize even one possible world in all its detail seems beyond the average person. Perhaps one might chalk this up as a harmless idealization, of the kind frequently appealed to in many measurement systems. Somewhat more worrying, however, is that convincing a subject that 'we had the power of the Almighty' would surely drastically alter her doxastic state prior to measuring it, as Jeffrey (1983, pp. 158–60) has noted. Likewise, when a subject is offered the choice of either w_3 or $(w_1, P; w_2)$, we must not suppose that her credence in P is in any way changed by the offer or this would ruin the measurement.

Interestingly, Ramsey himself objects to another proposed bet-based measurement system on the grounds that 'the proposal of the bet may inevitably alter [the subject's] state of opinion' (1931b, p. 172). Either Ramsey did not recognize that the same objection applies with greater force to his own account, or he believed that the worry could be addressed. Bradley (2001, pp. 285–8) suggests one way in which it might be addressed: rather than making the subject believe in our godlike powers, we simply ask her to choose amongst options *as if* they were genuinely available to her (with perhaps the added proviso that she ought not to change her credences in any relevant propositions). To the extent that such a request can be satisfied, this reconstrual of Ramsey's methodology may help to minimize any changes to subjects' credences prior to measurement.

In any case, we can now say precisely what Ramsey meant when he referred to the accuracy of expected utility theory in 'the sort of cases with which we are most concerned'. We are to limit our attention to conscious, deliberate and presumably reflective judgements of preference between worlds and worlds, gambles and gambles, and worlds and gambles. Plausibly, Ramsey would have also held that we are not to consider cases where the subject is intoxicated, or under any kind of substantial physical or emotional duress. Ramsey does not need to assume anything as strong as the truth of subjective expected utility theory *tout court*, nor even its approximate truth across a wide range of cases — he only needs to assume that it is accurate in this particular kind of case. It would not be misleading, therefore, to interpret Ramsey's use of $\succ$ as, roughly: $\alpha \succ \beta$ relative to an agent S just in case S prefers α to β after consciously deliberating on the matter, while neither under physical or emotional distress, nor under the influence

 © Elliott 2016

of any intoxicating substances. On this interpretation, Ramsey's assertion that expected utility theory is broadly accurate in 'the sort of cases with which we are most concerned' is essentially the claim that an ordinary agent's reflective preference ordering over worlds and gambles is what we would expect of an expected utility maximizer. Put like this, Ramsey's assumption appears quite plausible.

2.3 Defining an equal difference relation

Ramsey's first step has us empirically determine how the agent ranks outcomes and gambles. However, a simple preference ordering on outcomes and gambles only suffices for an ordinal scale representation of an agent's utilities. For Ramsey, this is unsatisfactory: 'There would be no meaning in the assertion that the difference in value between $[w_1]$ and $[w_2]$ was equal to that between $[w_3]$ and $[w_4]$' (1931b, p. 176). We usually suppose that it is meaningful to say such things as 'I desire α *much more* than β; more so than α over γ' and 'I prefer α over β to the same extent that I prefer γ over δ'. A merely ordinal representation is unable to capture this information, representing as it does only the relative positions of outcomes in the preference order. Thus Ramsey sets himself the task of characterizing an *equal difference* (in utilities) relation between pairs of outcomes wholly in terms of preferences over gambles. If he can do this, then on the basis of well-known results from the mathematical theory of measurement, he can construct a richer representation of our utilities.

Let us say that $(w_1, w_2) =^{d} (w_3, w_4)$ holds just in case the difference in value for the agent between w_1 and w_2 is equal to the difference in value between w_3 and w_4. Ramsey's goal of defining $=^{d}$ in terms of preferences over gambles then sets up a certain difficulty to be overcome. According to the assumed background theory, an agent's preferences over gambles are determined by two factors: their *utilities* and their *credences*. Whether an agent prefers $(w_1, P; w_2)$ to $(w_3, Q; w_4)$, for example, depends partly on the utilities that she attaches to the outcomes w_1, w_2, w_3, w_4, and partly on the credences she has with respect to P and Q. However, whether $(w_1, w_2) =^{d} (w_3, w_4)$ holds for that agent should depend *solely* on the utilities she attaches to w_1, w_2, w_3, w_4. In order to define $=^{d}$ in terms of preferences over gambles, then, Ramsey needs some way of factoring out any confounding influences, so that whether the agent prefers one of the relevant gambles to another depends *only* on the utilities attached to the outcomes involved.

Ramsey's central innovation here is to define, in terms of preference, what it is for a proposition to have probability ½, and then to

 © Elliott 2016

use this to define $=^{d}$. Let us suppose for simplicity that whether an agent prefers $(w_1, P; w_2)$ to $(w_3, Q; w_4)$ depends *only* on the utilities the agent has for w_1, w_2, w_3, w_4, and the credences she has for P and Q. More specifically, assume the following:

> *Naïve Expected Utility Theory*
> If *des* is a real-valued interval scale that measures the agent's utilities for outcomes (i.e., $des(w_1) \geq des(w_2)$ if and only if $w_1 \succcurlyeq w_2$), and *bel* is a probability function that represents the agent's credences, then $(w_1, P; w_2) \succcurlyeq (w_3, Q; w_4)$ if and only if $des(w_1).bel(P) + des(w_2).(1 - bel(P)) \geq des(w_3).bel(Q) + des(w_4).(1 - bel(Q))$.

We will note shortly that Ramsey did *not* assume Naïve Expected Utility Theory; but for now it suffices to explain the reasoning behind his definitions.

Suppose that the agent is indifferent between the following two gambles: $(w_1, P; w_2)$ and $(w_2, P; w_1)$. According to Naïve Expected Utility Theory, there are only two (not mutually exclusive) ways in which this might come about: either w_1 and w_2 have exactly the same utility for the agent, or the agent's credence in P is exactly 0.5. To rule out the former possibility, we consider a pair of gambles $(w_3, P; w_4)$ and $(w_4, P; w_3)$, where we know that the agent is not indifferent between w_3 and w_4. If we find that the agent is indifferent between $(w_3, P; w_4)$ and $(w_4, P; w_3)$, we will have established that $bel(P) = 0.5$. If the probability of P were any other way, then the agent would not have been indifferent between $(w_3, P; w_4)$ and $(w_4, P; w_3)$. For instance, if $w_3 \succ w_4$ and $bel(P) > 0.5$, then any minimally rational agent would prefer $(w_3, P; w_4)$ to $(w_4, P; w_3)$.

With this in place, we are then able to say that $(w_1, w_2) =^{d} (w_3, w_4)$ holds just in case $(w_1, P; w_4) \sim (w_2, P; w_3)$, where P has probability ½. The reasoning behind this is not immediately obvious, but goes as follows: from the assumption of Naïve Expected Utility Theory, we have that $(w_1, P; w_4) \sim (w_2, P; w_3)$ holds if and only if:

$$\begin{aligned} &des(w_1).bel(P) + des(w_4).(1 - bel(P)) = des(w_2).bel(P) \\ &\quad + des(w_3).(1 - bel(P)) \end{aligned}$$

We have also already established that $bel(P) = 0.5 = 1 - bel(P)$, so we can drop the constant factor leaving us with:

$$des(w_1) + des(w_4) = des(w_2) + des(w_3)$$

© Elliott 2016

which holds just in case:

$$des(w_1) - des(w_2) = des(w_3) - des(w_4)$$

This just states that the difference between w_1 and w_2 is equal to the difference between w_3 and w_4; so if $bel(P) = 0.5$, then $(w_1, P; w_4) \sim (w_2, P; w_3)$ if and only if $(w_1, w_2) =^{d} (w_3, w_4)$. Given Naïve Expected Utility Theory, by appealing to probability ½ propositions, we are able to use preferences to define an equal difference relation that depends solely on the differences in utility of the outcomes.

2.4 Locating ethically neutral propositions of probability ½

Before moving on to measuring utilities, however, Ramsey makes the following note:

> There is first a difficulty which must be dealt with; the propositions like P ... which are used as conditions in the [gambles] offered may be such that their truth or falsity is an object of desire to the subject. This will be found to complicate the problem, and we have to assume that there are propositions for which this is not the case, which we shall call ethically neutral. (1931b, p. 177)

This is the entirety of what Ramsey writes regarding his motivation for introducing ethically neutral propositions.

The idea is clear enough: Naïve Expected Utility Theory is mistaken, as it fails to take into account the utility that may attach to the gamble's condition, and how the condition might influence the valuation of outcomes. Assuming that w_1 is consistent with both P and $\neg P$, it is possible that an agent might attach a different value to $(w_1 \,\&\, P)$ than to $(w_1 \,\&\, \neg P)$. These are potentially quite different states of affairs with potentially different utilities, and the truth or falsity of P might make a great deal of difference to how the outcome w_1 is valued. As a rough example, suppose that in w_1 the agent has a puppy as a pet, while in w_2 she instead keeps a kitten, and let P be *puppies spread disease but kittens do not*; plausibly, $(w_1 \,\&\, P)$ will be valued quite differently than $(w_1 \,\&\, \neg P)$, and likewise for $(w_2 \,\&\, P)$ and $(w_2 \,\&\, \neg P)$.

Instead of Naïve Expected Utility Theory, and supposing w_1, w_2, w_3, and w_4 are each compatible with the relevant propositions, instead of Naïve Expected Utility Theory we should instead have a principle that implies that:

$$(w_1, P; w_2) \geq (w_3, Q; w_4)$$

 © Elliott 2016

if and only if:

$$des(w_1\&P).bel(P) + des(w_2\&\neg P).(1 - bel(P)) \geq des(w_3\&Q).bel(P) + des(w_4\&\neg Q).(1 - bel(Q))$$

It is easy to see that this fact invalidates the reasoning behind both the definition of a ½ probability proposition and the definition of $=^d$, for now we can no longer say that the agent's preferences between (w_1, P; w_2) and (w_3, Q; w_4) depend on their credences in P and Q and the utilities the agent has for w_1, w_2, w_3, and w_4. Rather, they actually depend on the agent's credences in P and Q and their utilities for (w_1 & P), (w_2 & $\neg P$), (w_3 & Q), and (w_4 & $\neg Q$).

Ramsey's solution to this difficulty is the ethically neutral proposition — a kind of proposition the truth or falsity of which is of absolutely no concern to the agent. Ramsey provides us with a problematic definition of the notion, which I will discuss further in §3.2. The apparent purpose of the notion, however, is that if P is ethically neutral, then the conjunction of P with an outcome has the same utility as the outcome itself, and similarly for the conjunction of $\neg P$ and the outcome. Setting aside Ramsey's own definition, we can say that P is *ethically neutral* whenever $w \sim (w \,\&\, P) \sim (w \,\&\, \neg P)$, for any outcome $w \in \boldsymbol{W}$ that is compatible with both P and $\neg P$.[3]

It is clear what the upshot of introducing ethically neutral propositions is supposed to be: so long as we are considering gambles conditional on ethically neutral propositions, we can without risk of error apply Naïve Expected Utility Theory. Thus Ramsey (1931b, pp. 177–8) happens upon the following two definitions:

> *Definition* 1: *Ethically neutral proposition of probability ½*
> P is an ethically neutral proposition of probability ½ if and only if P is ethically neutral, and for some w_1, $w_2 \in \boldsymbol{W}$, $\neg(w_1 \sim w_2)$, and $(w_1, P; w_2) \sim (w_2, P; w_1)$[4]

[3] If there are no outcomes compatible with both P and $\neg P$, then P is trivially ethically neutral by this definition. Ramsey likely would have required non-trivially ethically neutral propositions, though the issues here are complex. See §3.1 and §4.3 below for discussion relating to this issue.

[4] Ramsey states that 'We assume by an axiom that if this is true of any one pair [w_1], [w_2], it is true of all such pairs' (1931b, pp. 177–8). This axiom is never actually listed, and it does not seem to follow from any of (RAM 1)–(8) either separately or in conjunction.

 © Elliott 2016

And:

Definition 2: Equal difference relation
$(w_1, w_2) =^{d} (w_3, w_4)$ if and only if $(w_1, P; w_4) \sim (w_2, P; w_3)$, where P is an ethically neutral proposition of probability ½

2.5 Measuring utilities

At this point (1931b, pp. 178–9), Ramsey lists eight axioms, and states (but does not prove) that their satisfaction enables an appropriately rich representation of the agent's preferences. That is:

Ramsey's representation conjecture
If (RAM 1)–(8) hold, then there exists a real-valued function *des* on $\boldsymbol{W}$ such that for all $w_1, w_2, w_3, w_4 \in \boldsymbol{W}$, $des(w_1) - des(w_2) = des(w_3) - des(w_4)$ if and only if $(w_1, w_2) =^{d} (w_3, w_4)$; furthermore, *des* is unique up to positive linear transformation

We will not consider whether Ramsey's axioms successfully ensure the desired representation result, or how they might be fleshed out to do so if not—though see Bradley (2001) for relevant work in this regard. It is at least clear that something in the vicinity of Ramsey's axioms should suffice, however.

I here reproduce Ramsey's axioms, albeit with slightly improved notation. Let $\boldsymbol{P}$ be a set of propositions, $\boldsymbol{W}$ the set of outcomes, and $\boldsymbol{G}$ the set of gambles. There are multiple options for the formalization of $\boldsymbol{G}$. One might treat $\boldsymbol{G}$ as a set of functions from pairs of mutually exclusive and exhaustive propositions to $\boldsymbol{W}$ (as in Bradley 2001, p. 273), or simply take $\boldsymbol{G}$ to be a subset of $\boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{W}$ (as I do below, §4); the differences between these options need not concern us here. $\succ$ and $\sim$ are defined on $\boldsymbol{W} \cup \boldsymbol{G}$. The very first axiom is the most distinctive aspect of Ramsey's theorem:

(RAM 1) There is at least one ethically neutral proposition of probability ½

The importance of (RAM 1) for the rest of Ramsey's formal system should not be understated. Most of the axioms to follow are stated in terms of $=^{d}$, which is defined in terms of ethically neutral propositions. If (RAM 1) is false, those axioms will be in some cases false, in others trivial; in either case, the system as a whole collapses without this foundational assumption.

The next three axioms are each obviously necessary for Ramsey's desired representation result: for all $P, Q \in \boldsymbol{P}$, $w_1, w_2, w_3, w_4, w_5, w_6 \in \boldsymbol{W}$, $(w_1, P; w_2), (w_3, P; w_4) \in \boldsymbol{G}$, and $\alpha, \beta, \gamma \in \boldsymbol{W} \cup \boldsymbol{G}$,

 © Elliott 2016

(RAM 2) (i) If P, Q, are both ethically neutral propositions with probability ½, and $(w_1, P; w_2) \sim (w_3, P; w_4)$, then $(w_1, Q; w_2) \sim (w_3, Q; w_4)$; (ii) if $(w_1, w_2) =^{d} (w_3, w_4)$, then $w_1 \succ w_2$ iff $w_3 \succ w_4$, and $w_1 \sim w_2$ iff $w_3 \sim w_4$

(RAM 3) $\alpha \sim \beta$ and $\beta \sim \gamma$ only if $\alpha \sim \gamma$

(RAM 4) If $(w_1, w_2) =^{d} (w_3, w_4)$ and $(w_3, w_4) =^{d} (w_5, w_6)$, then $(w_1, w_2) =^{d} (w_5, w_6)$

The role of (RAM 2) is to ensure that the definition of $=^{d}$ is coherent. (RAM 3) simply states that $\sim$ is transitive, while (RAM 4) states that $=^{d}$ is transitive — an obvious requirement given the intended interpretation of $=^{d}$. Given the definition of the equal difference relation above, (RAM 4) is equivalent to the following:

> If $(w_1, P; w_4) \sim (w_2, P; w_3)$ and $(w_3, P'; w_6) \sim (w_4, P'; w_5)$, then $(w_1, P''; w_6) \sim (w_2, P''; w_5)$, where P, P', and P'' are ethically neutral propositions of probability ½

Together, (RAM 2)–(RAM 4) help to ensure that $=^{d}$, which holds between pairs of outcomes, mirrors the behaviour of the equals relation between the differences of pairs of real numbers.

The following two existential axioms are stated in terms of what Ramsey calls *values* — essentially, $\sim$-equivalence classes of outcomes. Formally:

> *Definition 3: The value of w*
> For every outcome $w \in \boldsymbol{W}$, let $\underline{w} = \{w' \in \boldsymbol{W}: w' \sim w\}$

The value of an outcome w, denoted $\underline{w}$, is thus the set of all outcomes in $\boldsymbol{W}$ with the same desirability as w. Ramsey's next two axioms are then:

(RAM 5) For all $\underline{w}_1$, $\underline{w}_2$, $\underline{w}_3$, there is exactly one $\underline{w}_4$ such that $(w_1, w_4) =^{d} (w_2, w_3)$

(RAM 6) For all $\underline{w}_1$, $\underline{w}_2$, there is exactly one $\underline{w}_3$ such that $(w_1, w_3) =^{d} (w_3, w_2)$

(RAM 5) implies that there is always at least one outcome w_4 such that the difference between w_1 and w_4 is equal to the difference between w_2 and w_3, for *any* choice of outcomes w_1, w_2, and w_3. In a manner of speaking, (RAM 6) says that for any pair of worlds w_1 and w_2, there is a third world w_3 with a utility which is exactly half-way between the

© Elliott 2016

utilities of w_1 and w_2. Given (RAM 1) (which implies the non-triviality of $\succ$ on $\boldsymbol{W}$), this entails a certain denseness to the agent's preference structure, and correspondingly, that $\boldsymbol{W}$ is infinite.

Finally, Ramsey lists two other axioms, which are not spelled out in any detail:

> (RAM 7) 'Axiom of continuity: — Any progression has a limit (ordinal)' (1931b, p. 179)
>
> (RAM 8) Archimedean axiom

What Ramsey intended for (RAM 7) is something of a mystery. One guess (cf. Sobel 1998 and Bradley 2001) would be that for every gamble $(w_1, P; w_2)$, there is an outcome w_3 such that $w_3 \sim (w_1, P; w_2)$. An axiom to this effect seems to be required to ensure that every real number can be mapped to at least one world's value.

Ramsey does not specify the character of (RAM 8), though it is easy to guess its role — like other so-called Archimedean axioms in various representation theorems, it is supposed to rule out any one outcome or gamble being incomparably better or worse than another. More specifically, (RAM 8) is intended to ensure that the numerical representation satisfies the Archimedean property of real numbers: for any positive number x, and any number y, there is an integer n such that $n + x \geq y$. This ensures that we do not require infinite desirability values to measure the agent's utilities. Were one to spell out (RAM 8), it is likely that it would need to look much like (ADS 5) of *Definition 13* below.

2.6 Measuring credences

Suppose that we have our function *des*. Ramsey then argues that:

> Having thus defined a way of measuring value we can now derive a way of measuring belief in general. If the option of $[w_2]$ for certain is indifferent with that of $[(w_1 \text{ if } P; w_3)]$, we can define the subject's degree of belief in P as the ratio of the difference between $[w_2]$ and $[w_3]$ to that between $[w_1]$ and $[w_3]$ ... This amounts roughly to defining the degree of belief in P by the odds at which the subject would bet on P, the bet being conducted in terms of differences of value as defined. (1931b, pp. 179–80)

In a footnote, Ramsey adds that '$[w_1]$ must include the truth of P, $[w_3]$ its falsity; P need no longer be ethically neutral' (1931b, p. 179, Fn. 1). We are led to the following definition:

 © Elliott 2016

Definition 4: Ramsey's credence function
For all contingent propositions P and outcomes w_1, w_2, w_3 such that w_1 implies P, w_3 implies $\neg P$, $\neg(w_1 \sim w_3)$, and $w_2 \sim (w_1, P; w_3)$, $bel(P) = (des(w_2) - des(w_3)) \,/\, (des(w_1) - des(w_3))$

This definition does not require P to be ethically neutral. Ramsey mistakenly states that it 'only applies to partial belief and does not include certain beliefs' (1931b, p. 180), though plausibly he meant that the definition does not apply if P is non-contingent. In this case, we simply stipulate that $bel(P) = 1$ if P is necessary, 0 if P is impossible. Note that, because ratios of differences are preserved across positive linear transformations of the *des* function, $bel(P)$ so-defined is unique.

The reasoning behind this final step is again left to the reader, though again it follows from Ramsey's background assumption of the descriptive adequacy of subjective expected utility theory. Note first of all that if w_1 entails P, then the conjunction of P and w_1 is just equivalent to w_1, so $des(w_1) = des(w_1 \,\&\, P)$. Ramsey clearly presupposes, then, that the desirability of a prospect $(w_1, P; w_2) \sim w_3$, where w_1 entails P and w_2 entails $\neg P$, is given by the following equality:

$$des((w_1, P; w_2)) = des(w_1).bel(P) + des(w_2).(1 - bel(P)) = des(w_3)$$

This is then rearranged to give us the definition of $bel(P)$ as above.

Ramsey does note that for his definition to work we require two more assumptions. The first of these is that the value of $bel(P)$ does not depend on the choice of worlds and gambles satisfying the stated conditions. In effect, this is to place restrictions directly upon *bel* after it has been defined in terms of preferences. (See *Condition 1*, §4.4 below.) The second assumption is that for any gamble $(w_1, P; w_2)$ we will always be able to find some world w_3 such that $w_3 \sim (w_1, P; w_2)$. Both of these assumptions are clearly required if *bel* is to be coherently defined for each contingent proposition.

Ramsey (1931b, pp. 180ff) goes on to define conditional probabilities using preferences over more complicated gambles, and he argues that *bel* satisfies the laws of probability, though I will not recapitulate that argument here: it is enough that Ramsey provides a credence function, *bel*: $\boldsymbol{P} \mapsto [0,1]$, that supposedly represents the agent's degrees of belief — after all, it combines with the agent's utilities for outcomes to determine their preference ordering for two-outcome gambles in more or less the manner we pre-theoretically expect degrees of belief

 © Elliott 2016

to. For our purposes, it is incidental whether *bel* satisfies the axioms of the probability calculus.

In summary, Ramsey's proposal is as follows. We begin with the background assumption of the empirical adequacy of subjective expected utility theory, and an empirically determined preference ordering on the set of outcomes and two-outcome gambles. Then, referring specifically to gambles conditional on an ethically neutral proposition of probability ½ that is assumed to exist (and determined through observation of the agent's preferences), we construct a function *des* that measures the agent's utilities. Having determined *des*, and given some further assumptions, we use *des* to determine the function *bel*. It is then argued that *bel* has certain attractive properties that lend support to its interpretation as a measure of the agent's degrees of belief.

3. Ethical neutrality

3.1 Why Ramsey introduced ethical neutrality

Ramsey was right to reject Naïve Expected Utility Theory. If the outcomes w_1 and w_2 are each compatible with both P and $\neg P$, then it is entirely possible that the agent values $(w_1 \,\&\, P)$ more (or less) than $(w_1 \,\&\, \neg P)$, and similarly for $(w_2 \,\&\, P)$ and $(w_2 \,\&\, \neg P)$. Any rational agent ought to take this into account when deliberating between gambles conditional on P with w_1 and w_2 as outcomes. For example, contrary to Naïve Expected Utility Theory, it is possible that the agent could be indifferent between $w_1 \sim w_2$ without thereby being indifferent between $(w_1, P; w_2)$ and $(w_2, P; w_1)$, if the truth or falsity of P makes a difference to how the agent values w_1 or w_2.

However, this point is conditional on w_1 and w_2 being each compatible with both P and $\neg P$. If instead we suppose that w_1 implies P, then $(w_1 \,\&\, P)$ is logically equivalent to w_1—and for Ramsey, if w_1 implies P, then the desirability of w_1 is just the desirability of $(w_1 \,\&\, P)$.[5] Ramsey's characterisation of the *bel* function relies on this very fact.[6]

[5] If w_1 implies $\neg P$, then $(w_1 \,\&\, P) = \varnothing$. It is unclear whether Ramsey intended his utility function to assign any value to $\varnothing$, as his definition of the credence function only requires the utility function to be defined on $\boldsymbol{W} \cup \boldsymbol{G}$, which does not contain $\varnothing$.

[6] While descriptively questionable, I will not discuss this idealizing assumption in any further detail. Such assumptions are standardly made in many descriptive decision theories and measurement systems—if it is a problem for Ramsey, then it is a problem for a great deal of the work in this area.

 © Elliott 2016

So, inasmuch as w_1 implies P and w_2 implies $\neg P$,

$$\begin{aligned} des((w_1, P; w_2)) &= des(w_1\&P).bel(P) + des(w_2\&\neg P).(1 - bel(P)) \\ &= des(w_1).bel(P) + des(w_2).(1 - bel(P)) \end{aligned}$$

Note that this holds regardless of whether P is ethically neutral. In other words, if w_1 implies P and w_2 implies $\neg P$, then we can apply Naïve Expected Utility Theory to the gamble $(w_1, P; w_2)$.

Interestingly, Ramsey originally describes his outcome set $\boldsymbol{W}$ as a set of possible worlds, and it is part of Ramsey's background theory that every world individually determines the truth or falsity of any proposition. In particular, Ramsey assumed a broadly Wittgensteinian logical atomism (though he believed it possible to reformulate his theorem without these commitments; see his 1931b, p. 177, fn. 1). Ramsey's discussion does not commit him to the entirety of Wittgenstein's theory. Instead, Ramsey assumes what Sobel (1998) calls a *thin* logical atomism, in the following sense: we are to suppose that there exists a class of atomic propositions such that no two worlds are exactly identical with respect to the truth of these propositions, every atomic proposition can be true or false entirely independently of any others, and for every world w and atomic proposition P, there is another world $w^\star$ that differs only with respect to the truth of P. Every possible world on this picture is determined by the set of atomic propositions true at that world. Even setting aside the assumption of logical atomism, on an orthodox conception of propositions as sets of worlds, for any given (determinate) proposition, a given world either is or is not a member of that proposition. Every world therefore determines either the truth or falsity of any proposition, and no possible world is compatible with both P and $\neg P$.

This leaves us with something of a puzzle: why did Ramsey alter his characterisation of the outcome set (as noted in §2.2)? It seems that if he instead limited his attention to gambles like $(w_1, P; w_2)$, where w_1 implies P and w_2 implies $\neg P$, then he could have used preferences over *these* to define $=^{d}$ without needing to introduce the notion of ethical neutrality. Let us call any gamble $(w_1, P; w_2)$ *impossible* if either w_1 implies $\neg P$ or w_2 implies P. A gamble is *possible* if and only if it is not impossible. If outcomes are possible worlds, then every possible gamble $(w_1, P; w_2)$ must be such that w_1 implies P and w_2 implies $\neg P$, and so $des(w_1) = des(w_1 \,\&\, P)$ and $des(w_2) = des(w_2 \,\&\, \neg P)$. We can therefore always apply Naïve Expected Utility Theory to possible gambles if the outcomes are worlds. So why did Ramsey not stick to his

 © Elliott 2016

original characterisation of outcomes as worlds, and simply use preferences over possible gambles to define $=^{d}$?

The answer to this question can be discovered by considering again how Ramsey defines ½ probability propositions. In particular, to determine whether P has a probability ½, we need to consider preferences over two gambles of the form $(w_1, P; w_2)$ and $(w_2, P; w_1)$. The definition Ramsey gives us *only* makes sense if the outcomes w_1 and w_2 are *not* possible worlds. If w_1 and w_2 are possible worlds, then at least one of the two gambles is impossible, and if either gamble is impossible, then the reasoning behind the assignment of probability ½ to the proposition P is no longer valid. Indeed, Ramsey recognized the difficulty here, and for this reason wrote that, at least for those outcomes w_1 and w_2 required for his definition of ½ probability propositions, '$[w_1]$ and $[w_2]$ must be supposed so far undefined as to be compatible with both P and $\neg P$' (1931b, p. 178, Fn. 1). Supposing for simplicity that P is atomic, we are presumably to take the outcomes w_1 and w_2 as *near-worlds*, which we can understand as propositions that are just shy of being maximally specific. Given thin logical atomism, for every world w and every atomic proposition P, there is a proposition that *nearly* uniquely identifies w but does not specify whether P is true. In Ramsey's framework, a near-world with respect to an atomic proposition P is a disjunction of two worlds w^{P} and $w^{\neg P}$ that are identical with respect to all of their atomic propositions except for P.

The answer to our puzzle, then, is that Ramsey's set of outcomes cannot quite be the set of possible worlds given his strategy for defining $=^{d}$. For the pair of possible gambles $(w_1, P; w_2)$ and $(w_2, P; w_1)$ referred to in *Definition 1*, neither w_1 nor w_2 can imply either P or $\neg P$. It follows for the reasons given, then, that we cannot in general apply Naïve Expected Utility Theory to such gambles, unless we appeal to ethically neutral propositions.[7]

[7] An anonymous referee suggests an alternative interpretation of what it is for w_1 and w_2 to be 'so far undefined as to be compatible with both P and $\neg P$': outcomes are in all cases possible worlds, while $(w_1, P; w_2)$ should be understood as (i) having outcome w_1 if P and w_1 implies P, or w_1' if P and w_1 implies $\neg P$, and (ii) likewise, w_2 if $\neg P$ and w_2 implies $\neg P$, or w_2' if $\neg P$ and w_2' implies P, where w_1' and w_2' are like w_1 and w_2 respectively in all atomic details except P. On this interpretation, Ramsey kept his original characterisation of outcomes as worlds, introducing considerable complexity into his understanding of a gamble instead.

It is clear why the notion of ethical neutrality is needed on this picture: if P is ethically neutral then $w_1 \sim w_1'$ and $w_2 \sim w_2'$, which would justify applying Naïve Expected Utility Theory to $(w_1, P; w_2)$ and $(w_2, P; w_1)$. On either interpretation, then, the main upshot of the present discussion is retained: Ramsey needed to introduce ethical neutrality largely as a result of his strategy for defining $=^{d}$.

 © Elliott 2016

3.2 Problems with ethical neutrality

In looking at whether the notion of ethical neutrality is viable, we ought first to start with Ramsey's own definition:

> *Definition 5: Ethical neutrality (Ramsey's original)*
> P is ethically neutral if and only if (i) if P is atomic, then $w^P \sim w^{\neg P}$, for all pairs of worlds w^P, $w^{\neg P}$ identical with respect to all their atomic propositions except for P, (ii) if P is non-atomic, then all of Ps atomic truth arguments are ethically neutral

So, an atomic proposition P is *ethically neutral* for an agent just in case any two possible worlds differing in their atomic propositions only in the truth of P are always equally valued by that agent, and ethical neutrality for non-atomic propositions is understood in terms of atomic propositions. Ramsey here demonstrates commitment to another aspect of Wittgensteinian atomism: every non-atomic proposition can be constructed from atomic propositions using truth-functional connectives.[8] We are able to locate such a proposition, if it exists, by considering the agent's preferences over worlds. As just noted, for some gambles $(w_1, P; w_2)$ and $(w_2, P; w_1)$, Ramsey requires that w_1 and w_2 are compatible with both P and $\neg P$. If we suppose for simplicity that P is atomic, then w_1 and w_2 are near-worlds with respect to P. It follows from Ramsey's definition then that $(w_1 \,\&\, P) \sim (w_1 \,\&\, \neg P)$ and $(w_2 \,\&\, P) \sim (w_2 \,\&\, \neg P)$. It does *not* yet follow that $(w_1 \,\&\, P) \sim (w_1) \sim (w_1 \,\&\, \neg P)$, which Ramsey also required. However, we can take this as an unstated background assumption: if $(w_1 \,\&\, P) \sim (w_1 \,\&\, \neg P)$, then $(w_1 \,\&\, P) \sim (w_1) \sim (w_1 \,\&\, \neg P)$.

Sobel (1998, p. 241) has argued that there are few or no ethically neutral propositions in this sense. Consider the proposition *there are an even number of hairs on Dan Quayle's head*. Sobel argues that this can be ethically neutral for 'almost no one':

> Though it is true that I do not care about Quayle's hair, there are worlds that differ regarding the truth of that proposition that, just because of that difference, differ in their values for me. I am thinking of worlds in which I have bet money on this proposition! The argument ... can be readdressed to atomic propositions, if such there be, to the conclusion that *no* atomic proposition is

[8] Interestingly, Ramsey's definition suggests that ethically neutral propositions can be logically equivalent to non-ethically neutral propositions. Suppose that P is ethically neutral, but Q is not. Then P is logically equivalent to $(P \,\&\, (Q \vee \neg Q))$, but $(P \,\&\, (Q \vee \neg Q))$ is not ethically neutral. Thanks to Rachael Briggs for noting this.

 © Elliott 2016

> Ramsey-ethically-neutral for any of us. Ramsey's existence axiom for ethically neutral atomic propositions, even if coherent, severely curtails the applicability of his theory. (Sobel 1998, p. 248, emphasis in original)

There seem to be two concerns here. The first appears to be something like the following: for any proposition whatsoever, we should be able to find a set of otherwise similar possible worlds where we have entered into a bet conditional on that proposition with desirable outcomes if things turn out one way, and undesirable outcomes if things turn out another way. Since we care about the outcomes of the bet, we will value the relevant worlds differently. However, this objection seems to have no hold given Ramsey's view: the relevant worlds are supposed to differ at the atomic level *only* with respect to the proposition in question. In all other respects — including, importantly, the payouts for any bets we may enter into — the worlds are supposed to be identical.

The second and more obvious worry is that Ramsey's conception of ethical neutrality requires the assumption of logical atomism for its cogency. Ramsey built his theory upon the assumption of logical atomism so that he could make sense of the idea of two worlds differing *only* with respect to a particular proposition. The notion is of little use to contemporary philosophers who by and large reject that aspect of Wittgenstein's view.

In his atomism-free reconstruction of Ramsey's theorem, Bradley (2001) supplies the following definition, intended to achieve the same purpose:

> *Definition 6: Ethical neutrality (atom-free)*
> P is ethically neutral if and only if for all propositions Q that are compatible with both P and $\neg P$, $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$

Tautological and impossible propositions will be trivially ethically neutral according to this definition. Clearly, however, we are interested only in non-trivially ethically neutral propositions. A common suggestion is that propositions such as *the tossed coin will land heads* constitute ethically neutral propositions of probability ½. Part of the reason why we use coin tosses occasionally when making decisions is because we have no intrinsic interest in whether the coin lands heads or tails. If Q is something like *there are dogs*, and P is *the tossed coin will land heads*, then it seems plausible that $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$.

 © Elliott 2016

However, there are strong reasons to think that *no* contingent propositions will be ethically neutral in the sense of *Definition 6*, for any minimally rational subject. Let P be *the tossed coin will land heads*, and take Q to be the proposition:

> (*the tossed coin will land heads* & *I receive $100,000*) *or* (*the tossed coin will not land heads* & *I get kicked in the shins*)

Q is obviously compatible with both P and $\neg P$. However, $(P \,\&\, Q)$ is equivalent to *the tossed coin will land heads* & *I receive $100,000* while $(\neg P \,\&\, Q)$ is equivalent to *the tossed coin will not land heads* & *I get kicked in the shins*. But for some very strange preference orderings, it is certainly not the case that $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$. The point here generalizes easily; there are no non-trivially ethically neutral propositions in this sense. Note that the issue here is not that no contingent proposition satisfies the definition *exactly*, while there may nevertheless be some propositions which *approximate* ethical neutrality. Rather, the upshot is that no proposition even comes *close* to satisfying the requirements of ethical neutrality. We will always be able to find countlessly many propositions Q that falsify the indifference requirements.[9]

A refinement of *Definition 6* might be useful here. Instead of requiring $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$ for all *propositions* Q compatible with both P and $\neg P$, Ramsey only requires the following:

> *Definition 7: Ethical neutrality (atom-free, refined)*
> P is ethically neutral if and only if $w \sim (w \,\&\, P) \sim (w \,\&\, \neg P)$, for any outcome $w \in \boldsymbol{W}$ that is compatible with both P and $\neg P$

If there are no outcomes compatible with both P and $\neg P$, then P is trivially ethically neutral by this definition. Again, we can set such propositions aside; we are interested in non-trivially ethically neutral propositions. (See §4.3 for more discussion.) If Q is not in the outcome set $\boldsymbol{W}$, then there are no relevant gambles with Q as an outcome and we do not need to concern ourselves over whether $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$. If we assume that there are far fewer propositions in $\boldsymbol{W}$ than in $\boldsymbol{P}$, then the foregoing objection is blocked. This will certainly be true if the outcomes in $\boldsymbol{W}$ are highly specific, as is the case in Ramsey's system.

With that said, it is still not obvious that any non-trivially ethically neutral propositions exist even in this weaker sense. Why should we

[9] Thanks to Rachael Briggs and Alan Hájek for these points and related discussion.

 © Elliott 2016

suppose that there are *any* propositions P such that (non-trivially), $w \sim (w \,\&\, P) \sim (w \,\&\, \neg P)$ for all $w \in \boldsymbol{W}$ compatible with P and $\neg P$? Indeed, without knowing the exact nature of the outcome space $\boldsymbol{W}$, we cannot even know whether there *are* any outcomes compatible with both P and $\neg P$, for an arbitrarily chosen proposition P. Ramsey explicitly stipulates that there must be at least one pair of outcomes compatible with *some* ½ probability ethically neutral proposition and its negation — but this stipulation is meaningless inasmuch as we do not already know what proposition that may be. Unfortunately, Ramsey's discussion leaves the nature of $\boldsymbol{W}$ quite vague, making the matter impossible to judge.

We can circumvent this concern by stipulating that $\boldsymbol{W}$ contains, for each of a very wide range of propositions in $\boldsymbol{P}$, highly specific outcomes that are undefined with respect to that proposition. But even then, Ramsey gives us little reason to suppose that ethically neutral propositions exist relative to a given agent's preference ordering — still less to suppose that there are any such propositions that satisfy *Definition 1*. (RAM 1) clearly cannot be defended as a condition of rationality, and it does not follow from Ramsey's background assumption of the descriptive adequacy of expected utility theory. Ramsey's aim in the first instance was to develop a procedure for the measurement of degrees of belief, so unlike other intended uses for decision theoretic representation theorems he did not require his axioms to be constraints of practical rationality; nevertheless, if the process is to be viable then it ought at least to be *applicable*. It may not be impossible for a rational agent to satisfy the axiom, but we still require good reasons to believe that most do — yet no reasons are forthcoming.

A related issue regards Ramsey's proto-functionalist attempt to define degrees of belief in terms of his measurement procedure: a definition of credences which relies on a dubitable and unjustified existential assumption is, at best, of very limited interest. Are we to suppose that agents who falsify (RAM 1) do not have degrees of belief? Ultimately, given his reliance upon ethically neutral propositions, Ramsey's system was not sufficient to establish the main upshot of 'Truth and Probability', namely, that the laws of probability provide for us the logic of partial belief. Even if it is understood in terms of *Definition 7*, (RAM 1) is a very shaky foundation for a measurement procedure, and still worse for a definition of credences.

Many expected utility representation theorems developed since Ramsey's original work have also made use of ethically neutral

 © Elliott 2016

propositions, either explicitly or implicitly. Davidson and Suppes (1956) develop a representation theorem similar to Ramsey's wherein they explicitly characterize and axiomatize the existence of ethically neutral propositions. Others make implicit appeal to ethically neutral propositions, in the sense that they figure in the intended interpretation of the formal system, rather than being formalized directly. In this capacity we find ethical neutrality in the theorems of Davidson, Siegel, and Suppes (1957), Debreu (1959), Suppes (1956), and Fishburne (1967). Each of these works appears to require an understanding of ethical neutrality in something like the senses of *Definition 6* or *7* (each for essentially the same reason that Ramsey required the notion), and thus they inherit the problems associated with those definitions. We need a firmer foundation for the measurement of credences.

4. Refining Ramsey's system

4.1 Preliminaries

Ramsey's motivation for introducing the idea of ethical neutrality arises ultimately from his strategy for defining ½ probability propositions and $=^{d}$. But we are not forced to use Ramsey's definitions. It is possible to avoid introducing ethical neutrality in any of the senses defined so far, if we can develop alternative definitions.

One obvious possibility would be to introduce a second qualitative relation into our formal system, à la Joyce (1999). Let $=^{b}$ be a *qualitative probability relation* defined on the set of propositions: $P =^{b} Q$ if and only if the agent judges P to be exactly as likely as Q. Now we might say that a proposition P has probability ½ just in case $P =^{b} \neg P$. Developing this line of thought would lead us to a two-primitive (*preference* and *qualitative probability*) representation theorem, though it is also likely to render much of Ramsey's approach redundant. For instance, if the agent has a sufficiently rational qualitative probability ranking $\geq^{b}$ on the set of propositions, then we can apply something like de Finetti's (1931) representation theorem to arrive at a cardinal representation of her credences without needing to consider her preferences over gambles at all.

Let us suppose, then, that we have no straightforward empirical access to agents' qualitative probability rankings. We are then in Ramsey's original position, aiming to develop a system for the

 © Elliott 2016

measurement of credences where we only have empirical access to agents' preferences over outcomes and gambles.

Another possibility would be to construct a *relative* notion of ethical neutrality. Ramsey's definition of $=^d$ in terms of gambles does not need ethical neutrality *tout court*; instead, the conditioning proposition need only be ethically neutral with respect to the outcomes involved in the relevant gambles. That is, for Ramsey, $(w_1, w_2) =^d (w_3, w_4)$ if and only if $(w_1, P; w_4) \sim (w_2, P; w_3)$, where P is an ethically neutral proposition of probability ½, but here it is enough if P is *neutral with respect to* w_1, w_2, w_3, and w_4. Sobel (1998, pp. 268–9) suggests a modification to Ramsey's system along these lines, which forces the replacement of (RAM 1) with:

(RAM 1′) For every quartet of worlds there is a proposition believed to degree ½ that is ethically neutral with respect to each world in the quartet

The difficulty here then becomes spelling out the notion of relative ethical neutrality in such a manner as to render (RAM 1′) plausible. Sobel's own suggested definition also presupposes logical atomism, as it was intended to fit with the rest of Ramsey's system. A more straightforward strategy, however, is to bypass the issue of ethical neutrality altogether; we can have an essentially Ramseyian representation theorem without needing to first construct a notion of ethical neutrality at all, rendering attempts to define the notion pointless.

Before we move on, I will note two assumptions that underlie the rest of the discussion. First, I will follow Ramsey's initial suggestion and take the space of outcomes, $\boldsymbol{W}$, to be the set of all possible worlds.[10] This assumption is for simplicity only; nothing of importance to the formal result is altered if we take outcomes to be propositions. (See the discussion in §4.3.) Secondly, and far more importantly, I will limit my attention to *possible* gambles only. I am doubtful that it makes much sense to assert that an agent can have interesting preferences with respect to impossible gambles, and in any case their inclusion comes with a great deal of added complexity. Preferences over impossible gambles play no role in deriving the

[10] I leave open the kind of modality being employed here. One might take $\boldsymbol{W}$ to be the set of metaphysically possible worlds, though this may lead to some counterintuitive results. My preference is to construe $\boldsymbol{W}$ as the set of epistemically possible worlds; in the sense of Chalmers (2011).

 © Elliott 2016

forthcoming representation result. I axiomatize this limitation in (GRS 1) below.

The key idea of the theorem below is that, while Ramsey used the same *outcomes* in two different gambles to define what it is for a proposition to have probability ½, this is unnecessary—it is enough if we instead use outcomes with exactly the same *desirability*. Suppose that $w_1 \sim w'_1$, $w_2 \sim w'_2$, and $\neg(w_1 \sim w_2)$, but:

$$(w_1, P; w_2) \sim (w'_2, P; w'_1)$$

We assume that each of these two gambles is possible; that is, w_1 and w'_2 each imply P, and w_2 and w'_1 each imply $\neg P$. Given the Ramseyian background assumption, this is possible only if:

$$des(w_1).bel(P) + des(w_2).(1 - bel(P)) = des(w'_2).bel(P) + des(w'_1).(1 - bel(P))$$

Since $w_1 \sim w'_1$ and $w_2 \sim w'_2$, we know that $des(w_1) = des(w'_1) = x$ and $des(w_2) = des(w'_2) = y$; and because $\neg(w_1 \sim w_2)$, we know that $x \neq y$. Let $bel(P) = z$. We are left with:

$$xz + y(1 - z) = yz + x(1 - z)$$

Regardless of the specific values of x and y, this is possible only if $z = (1 - z)$; thus, $bel(P) = 0.5$. This is essentially the same reasoning that Ramsey used in developing his definition of ½ probability propositions, with the only difference being that we have appealed to equally valued outcomes rather than using the same outcomes for the two gambles. Since $(w_1, P; w_2)$ and $(w'_2, P; w'_1)$ are both possible gambles with worlds as outcomes, we can apply Naïve Expected Utility Theory. There is no reason to require that P is ethically neutral. A set Π of ½ probability propositions is characterized precisely in *Definition 8* below.

Making this modification, in light of the two assumptions noted above, forces a number of changes to Ramsey's axiomatisation. There are two particularly important changes that I will note here, before laying out the main theorem in full. First, we can no longer employ Ramsey's definition of $=^d$. (Instead of defining $=^d$, I will instead define $\geq^d$.) However, here we can employ the same trick: there is no reason why $(w_1, w_2) \geq^d (w_3, w_4)$ must be defined using the worlds w_1, w_2, w_3, and w_4. It is enough if we use worlds with exactly the same desirability. Furthermore, there is no reason why we need to

 © Elliott 2016

use the same ½ probability proposition in both gambles. Instead, we can say that $(w_1, w_2) \geq^{d} (w_3, w_4)$ holds just in case, if $w_1 \sim w'_1$, $w_2 \sim w'_2$, $w_3 \sim w'_3$, $w_4 \sim w'_4$, and both $(w'_1, P; w'_4)$ and $(w'_2, P'; w'_3)$ are possible gambles where P and P' both have a probability of ½, then:

$$(w'_1, P; w'_4) \geq (w'_2, P'; w'_3)$$

The reasoning behind this is essentially identical to that outlined in §2.3. Note that w_1 need not be distinct from w'_1; the only thing we require is that w'_1 implies P (and similarly for the other outcomes, *mutatis mutandis*). $\geq^{d}$ is characterized precisely in *Definition 9* below.

Secondly, we need to ensure that there are enough worlds for this definition of $\geq^{d}$ to work. That is, we need to assume that we will always be able to find the required gambles $(w'_1, P; w'_4)$ and $(w'_2, P'; w'_3)$. This is not obviously going to be the case, if we are limiting our attention to possible gambles only. In effect, we need to assume that for every pair of worlds w_1 and w_2, there will always exist at least one proposition P of probability ½ such that for some worlds $w'_1 \sim w_1$ and $w'_2 \sim w_2$, w'_1 implies P and w'_2 implies $\neg P$. This is a reasonably strong assumption. It requires that every value contains multiple members, and that at least two of these members will disagree with respect to P for some P of probability ½. In effect, this assumption replaces Ramsey's axiom (RAM 1). It is formalized as (GRS 2) below.

4.2 Main representation theorem

In this section I reproduce the main formal result of the paper: a representation theorem for the construction of an interval scale *des* on the set of gambles $\boldsymbol{G}$ and outcomes $\boldsymbol{W}$ such that for all α, $\beta \in \boldsymbol{W} \cup \boldsymbol{G}$ and all $w_1, w_2, w_3, w_4 \in \boldsymbol{W}$,

$\alpha \succcurlyeq \beta$ if and only if $des(\alpha) \geq des(\beta)$
$(w_1, w_2) \geq^{d} (w_3, w_4)$ if and only if $des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$

In §4.4 I will introduce some further conditions that suffice for the construction of a credence function *bel*. Assume, in all that follows:

$\boldsymbol{W}$ is a non-empty set of possible worlds
$\boldsymbol{P}$ is an algebra of sets on $\boldsymbol{W}$ (i.e., a set of propositions)
$\boldsymbol{G} \subseteq \boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{W}$
$\succcurlyeq$ is a binary relation defined on $\boldsymbol{W} \cup \boldsymbol{G}$; $\succ$ and $\sim$ are defined in the usual way

 © Elliott 2016

Note that $\boldsymbol{P}$ need not contain all sets of worlds. For instance, the system to be developed is consistent with supposing that $\boldsymbol{P}=\{\varnothing, \boldsymbol{W}, P, \neg P\}$, where P (or $\neg P$) has probability ½. Given (GRS 1), $\boldsymbol{G}$ will be restricted to all and only possible gambles.[11]

In the sequel, I have adopted the notational convention that sameness of subscript for outcomes implies sameness of desirability (but the reverse need not hold). For instance, it should be assumed in all that follows that w_1' and w_1'' each refer to outcomes with the same desirability as w_1 (i.e., $w_1 \sim w_1'$ and $w_1' \sim w_1''$). It should not be assumed, however, that either w_1' or w_1'' is necessarily distinct from w_1. Likewise, $(w_1, P; w_2)$ should be understood as a variable for gambles with outcome w_1 if P, w_2 otherwise; and $(w_1', P; w_2')$ for gambles conditional on P with outcomes *equal in value* to w_1 and w_2. Again, the pair $(w_1, P; w_2)$ and $(w_1', P'; w_2')$ need not be distinct.

We first define the set of ½ probability propositions:

Definition 8: *The set of ½ probability propositions*, Π
$\Pi = \{P \in \boldsymbol{P}$: there are $w_1, w_2 \in \boldsymbol{W}$ such that $(w_1, P; w_2), (w_2', P; w_1') \in \boldsymbol{G}$, $\neg(w_1 \sim w_2)$, and $(w_1, P; w_2) \sim (w_2', P; w_1')\}$

Henceforth, I will use π, π', and so on, to designate propositions within Π. It should not be assumed that $\pi \neq \pi'$. Given this, I will use $(w_1, \pi; w_2)$ specifically for gambles conditional on *some* π in Π (with outcomes w_1 and w_2).

We can now define $\geq^{d}$:

Definition 9: *The difference between* w_1, w_2 *is at least as much as between* w_3, w_4
$(w_1, w_2) \geq^{d} (w_3, w_4)$ if and only if $(w_1', \pi; w_4') \succcurlyeq (w_2', \pi'; w_3')$ for all $(w_1', \pi; w_4'), (w_2', \pi'; w_3') \in \boldsymbol{G}$

For the purposes of characterizing the Archimedean axiom, we will also need to define a strictly bounded standard sequence. We can break this notion down into two concepts:

Definition 10: *Standard sequence*
$w_1, w_2, \ldots, w_i, \ldots$ is a standard sequence if and only if (i) for all $(w_2', \pi; w_1'), (w_1'', \pi'; w_1''') \in \boldsymbol{G}$, $\neg((w_2', \pi; w_1') \sim (w_1'', \pi'; w_1'''))$, and (ii) for every w_i, w_{i+1} in the sequence, $(w_{i+1}', \pi; w_2') \sim (w_1', \pi'; w_i')$ for all $(w_{i+1}', \pi; w_2'), (w_1', \pi'; w_i') \in \boldsymbol{G}$

[11] Note that $(w_1, P; w_2)$ and $(w_2, \neg P; w_1)$ are distinct elements of $\boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{W}$; here I leave it undetermined whether they represent distinct objects of preference. See §4.4, *Theorem 4*.

 © Elliott 2016

In light of the axioms to be characterized shortly, it will turn out that w_1, $w_2, \ldots, w_i, \ldots$ is a standard sequence just in case (i) $(w_2, w_1) \neq^d (w_1, w_1)$ and (ii) $(w_{i+1}, w_i) =^d (w_2, w_1)$ for all w_i, w_{i+1} in the sequence. So, for instance, the sequence w_1, w_2, w_3, w_4 is a standard sequence just in case:

$$(w_2, w_1) \neq^d (w_1, w_1) \text{ and } (w_4, w_3) =^d (w_3, w_2) =^d (w_2, w_1)$$

The idea, of course, is that the (nonzero) difference in desirability between any two adjacent members in the sequence is always equal to the difference in desirability between any other two adjacent members.

Definition 11: *Strictly bounded standard sequence*
$w_1, w_2, \ldots, w_i, \ldots$ is a strictly bounded standard sequence if and only if $w_1, w_2, \ldots, w_i, \ldots$ is a standard sequence and there exists w_a, $w_b \in \boldsymbol{W}$ such that for all w_i in the sequence, $(w'_a, \pi; w'_i) \succ (w'_1, \pi'; w'_b)$ and $(w''_i, \pi''; w''_b) \succ (w''_a, \pi'''; w''_1)$, for all $(w'_a, \pi; w'_i)$, $(w'_1, \pi'; w'_b)$, $(w''_i, \pi''; w''_b)$, $(w''_a, \pi'''; w''_1) \in \boldsymbol{G}$

In other words, any standard sequence $w_1, w_2, \ldots, w_i, \ldots$ is strictly bounded if there are w_a, $w_b \in \boldsymbol{W}$ such that for any w_i in the sequence, $(w_a, w_b) >^d (w_i, w_1) >^d (w_b, w_a)$. Essentially, regardless of the size of the interval between w_i and w_1, we can find outcomes in $\boldsymbol{W}$ that are spaced even further apart.

The coherence of the foregoing definitions will be ensured by the axioms (GRS 1)–(9), which we can now specify:

Definition 12: *Generalized Ramsey structure*
$<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a *generalized Ramsey structure* if and only if $\boldsymbol{W}$ is non-empty, $\boldsymbol{P}$ is an algebra of sets on $\boldsymbol{W}$, $\boldsymbol{G} \subseteq \boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{W}$, $\succcurlyeq$ is a binary relation on $\boldsymbol{W} \cup \boldsymbol{G}$, and for all $w_1, w_2 \in \boldsymbol{W}$, all sequences $w_1, w_2, \ldots, w_i, \ldots \in \boldsymbol{W}$, all $P \in \boldsymbol{P}$, and all $(w_1, P; w_2)$, $(w_1, \pi; w_2)$, $(w'_2, \pi'; w'_1)$, $(w_1, \pi; w_4)$, $(w_2, \pi'; w_3)$, $(w_3, \pi''; w_6)$, $(w_4, \pi'''; w_5) \in \boldsymbol{G}$, the following nine conditions hold:

(GRS 1) $(w_1, P; w_2) \in \boldsymbol{G}$ iff $w_1, w_2 \in \boldsymbol{W}$, $P \in \boldsymbol{P}$, and $w_1 \in P$, $w_2 \in \neg P$

(GRS 2) For every pair of worlds $w_1, w_2 \in \boldsymbol{W}$, there exists a $\pi \in \Pi$ such that for some $w'_1, w'_2 \in \boldsymbol{W}$, $w'_1 \in \pi$ and $w'_2 \in \neg\pi$

(GRS 3) $< \boldsymbol{W} \cup \boldsymbol{G}, \succcurlyeq>$ is a weak order

(GRS 4) If $(w_1, \pi; w_2)$, $(w'_2, \pi'; w'_1) \in \boldsymbol{G}$, then $(w_1, \pi; w_2) \sim (w'_2, \pi'; w'_1)$

 © Elliott 2016

(GRS 5) If $(w_1, \pi; w_4) \succcurlyeq (w_2, \pi'; w_3)$ and $(w_3, \pi''; w_6) \succcurlyeq (w_4, \pi'''; w_5)$, then, for all $(w'_1, \pi^*; w'_6)$, $(w'_2, \pi^+; w'_5) \in \boldsymbol{G}$, $(w'_1, \pi^*; w'_6) \succcurlyeq (w'_2, \pi^+; w'_5)$

(GRS 6) For every triple $w_1, w_2, w_3 \in \boldsymbol{W}$, there is a $w_4 \in \boldsymbol{W}$ such that for some $(w'_1, \pi; w'_3)$, $(w_4, \pi'; w'_2) \in \boldsymbol{G}$, $(w'_1, \pi; w'_3) \sim (w_4, \pi'; w'_2)$

(GRS 7) If $w_1, w_2, \ldots, w_i, \ldots$ is a strictly bounded standard sequence, then it is finite

(GRS 8) $w_1 \succcurlyeq w_2$ iff for all $(w'_1, P; w'_2) \in \boldsymbol{G}$, $w_1 \succcurlyeq (w'_1, P; w'_2) \succcurlyeq w_2$

(GRS 9) For each $(w_1, P; w_2) \in \boldsymbol{G}$, there is a $w_3 \in \boldsymbol{W}$ such that $(w_1, P; w_2) \sim w_3$

We can now state the main representation theorem:

Theorem 1: *Generalized Ramseyian utility measurement*
If $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure then there is a function *des*: $\boldsymbol{W} \cup \boldsymbol{G} \mapsto \mathbb{R}$ such that for all α, β in $\boldsymbol{W} \cup \boldsymbol{G}$ and all $w_1, w_2, w_3, w_4 \in \boldsymbol{W}$,

(a) $\alpha \succcurlyeq \beta$ if and only if $des(\alpha) \geq des(\beta)$

(b) $(w_1, w_2) \geq^{d} (w_3, w_4)$ if and only if $des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$

Furthermore, *des* is unique up to positive linear transformation

A proof is provided in the appendix. We now turn to a discussion of the axioms (GRS 1)–(9) before turning to how to get from generalized Ramsey structures to credence functions.

4.3 Generalized Ramsey structures

The strategy for proving *Theorem 1* is closely connected to Ramsey's process; given the agent's preferences over outcomes and two-outcome gambles, we first determine the relation $\geq^{d}$ between pairs of outcomes and on that basis construct an interval scale measurement of the agent's preferences. The most important step here is establishing that if $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure, then $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ is an *algebraic difference structure*:

Definition 13: *Algebraic difference structure*
$<\boldsymbol{A} \times \boldsymbol{A}, \succcurlyeq^*>$ is an *algebraic difference structure* if and only if $\boldsymbol{A}$ is non-empty, $\succcurlyeq^*$ is a binary relation on $\boldsymbol{A} \times \boldsymbol{A}$, and for all $a_1, a_2, a_3,$

 © Elliott 2016

$a_4, a'_1, a'_2, a'_3 \in \boldsymbol{A}$, and all sequences $a_1, a_2, \ldots, a_i, \ldots \in \boldsymbol{A}$, the following five conditions hold:

(ADS 1) $<\boldsymbol{A} \times \boldsymbol{A}, \succcurlyeq^*>$ is a weak order

(ADS 2) If $(a_1, a_2) \succcurlyeq^* (a_3, a_4)$, then $(a_4, a_3) \succcurlyeq^* (a_2, a_1)$

(ADS 3) If $(a_1, a_2) \succcurlyeq^* (a_4, a_5)$ and $(a_2, a_3) \succcurlyeq^* (a_5, a_6)$, then $(a_1, a_3) \succcurlyeq^* (a_4, a_6)$

(ADS 4) If $(a_1, a_2) \succcurlyeq^* (a_3, a_4) \succcurlyeq^* (a_1, a_1)$, then there exist $a_5, a_6 \in \boldsymbol{A}$ such that $(a_1, a_5) \sim^* (a_3, a_4) \sim^* (a_6, a_2)$

(ADS 5) If $a_1, a_2, \ldots, a_i, \ldots$ is such that $(a_{i+1}, a_i) \sim^* (a_2, a_1)$ for every a_i, a_{i+1} in the sequence, $\neg((a_2, a_1) \sim^* (a_1, a_1))$, and there exist $a', a'' \in \boldsymbol{A}$ such that $(a', a'') \succ^* (a_i, a_1) \succ^* (a'', a')$ for all a_i in the sequence, then it is finite

This allows us to invoke the following theorem:

> *Theorem 2: Algebraic difference measurement*
> If $<\boldsymbol{A} \times \boldsymbol{A}, \succcurlyeq^*>$ is an algebraic difference structure, then there exists a real-valued function f on $\boldsymbol{A}$ such that, for all $a_1, a_2, a_3, a_4 \in \boldsymbol{A}$, $(a_1, a_2) \succcurlyeq^* (a_3, a_4)$ if and only if $f(a_1) - f(a_2) \geq f(a_3) - f(a_4)$; furthermore, f is unique up to positive linear transformation

For a proof of *Theorem 2*, see Krantz *et al* (1971, Ch. 4).

Though none of the axioms (GRS 1)–(9) is identical to any of Ramsey's listed axioms, the majority of them bear a close resemblance to the axioms and assumptions mentioned by Ramsey in his paper. The differences are largely due to the assumption that the set of gambles includes only possible gambles, and the altered definition of ½ probability propositions. In this section I will discuss the axioms individually. It is worth noting first that none of the axioms is intended to be independently plausible *qua* norms of practical rationality, though at least a few may seem to have this status.[12] As with Ramsey's axiom system, the goal here is to establish conditions for the possibility of measurement under the assumption of the broad

[12] Plausibly, (GRS 3), (4), and (9) are constraints of practical rationality. I am inclined to take (GRS 5) as a rationality constraint, though this is difficult to justify without presupposing the representation result. The status of the Archimedean axiom (GRS 7) is unclear, though representation theorems that forego an Archimedean axiom can be developed, e.g., Bartha (2007). The existential axioms (GRS 2), (6) and (8) are not plausibly rationality constraints. (GRS 1) is a purely structural axiom, and places no constraints on any agent whether ideal or not.

 © Elliott 2016

descriptive adequacy of something like expected utility theory — we are not directly interested in establishing foundations for a prescriptive decision theory.

The purely structural axiom (GRS 1) does not correspond to any of Ramsey's axioms or any of the further assumptions he mentions, and his discussion is too sparse to know with any certainty whether he implicitly assumed anything like it. The move from worlds to near-worlds as outcomes suggests that he desired to avoid impossible gambles. However, given how Ramsey proposed to define $=^{d}$, without changes elsewhere in his system he also required either that we have preferences over impossible gambles, or that every outcome in $\boldsymbol{W}$ was compatible with both the truth and the falsity of some ethically neutral proposition. The argument for this proceeds by first noting that if $des(w_1) - des(w_2) = des(w_3) - des(w_4)$, then it should be the case that $(w_1, w_2) =^{d} (w_3, w_4)$. Suppose that $w_1 \sim w_1'$, so $des(w_1) - des(w_1') = des(w_1') - des(w_1)$. From *Definition 2*, we know that $(w_1, w_1') =^{d} (w_1', w_1)$ is only defined if the agent has preferences over some pair of gambles of the form $(w_1, P; w_1)$ and $(w_1', P; w_1')$ for some ethically neutral P of probability ½. It follows that either w_1 is compatible with P and $\neg P$, and similarly for w_1', or at least one of these two gambles is impossible.

We might suppose that Ramsey was happy to deal with preferences over impossible gambles. This would have forced him to assume that there is an interesting difference between two impossible propositions $(w_1 \,\&\, P)$ and $(w_2 \,\&\, P)$, where both w_1 and w_2 entail $\neg P$ but $\neg(w_1 \sim w_2)$. For suppose that Ramsey had only one impossible proposition, $\varnothing$. Then $des(w_1 \,\&\, P) = des(w_2 \,\&\, P) = des(\varnothing)$, but $des(w_1) \neq des(w_2)$. For whatever value we take $des(\varnothing)$ to have, it is clear that this will lead to problems. For instance, suppose that $des(\varnothing) \neq des(w_1)$; w_1 and w_2 each imply P; w_3 implies $\neg P$; and $des(w_1) = x$, $des(w_2) = des(w_3) = y$. We require that $(w_1, w_2) =^{d} (w_1, w_3)$, for obviously $x - y = x - y$. However, the justification for the definition of $=^{d}$ in terms of preferences fails under these conditions:

$$(w_1, w_2) =^{d} (w_1, w_3) \text{ if and only if } (w_1, P; w_3) \sim (w_2, P; w_1)$$

But this holds just in case:

$$des(w_1\&P).bel(P) + des(w_3\&\neg P).(1 - bel(P)) = des(w_2\&P).bel(P) + des(w_1\&\neg P).(1 - bel(P))$$

© Elliott 2016

Supposing P is ethically neutral and has probability ½, this equals:

$$0.5x + 0.5y = 0.5y + 0.5des(\varnothing)$$

It follows that $des(\varnothing) = x = des(w_1)$, which contradicts our initial assumption.

The only consistent way that Ramsey could have involved impossible gambles in his system would have been to treat different impossible propositions as different objects of desire. Perhaps an appeal to impossible worlds would suffice for this purpose: the impossible prospect of being a married bachelor might be desired to a greater degree than the prospect of being a square circle.

I do not want to judge whether Ramsey intended to countenance impossible gambles; for my own part, restricting our attention to possible gambles seems the better option. It is not obvious how we ought to treat preferences with respect to impossible propositions, if indeed there is more than one such proposition. For instance, it is implicit in Ramsey's system that if $w_1 \succcurlyeq w_2$, then $w_1 \succcurlyeq (w_1, P; w_2) \succcurlyeq w_2$. Without some such assumption he is unable to show that the function *bel* is a credence function. (See the proof in the appendix.) However, if we know that $(w_1 \,\&\, P)$ is impossible, it is not obvious why this should be the case: it seems at least as plausible that $w_2 \succcurlyeq (w_1, P; w_2)$ in this case, as we know we are not going to receive w_1 in the event that P and there is only a potentially very small probability $bel(\neg P) \leq 1$ of receiving w_2. Because it explicitly avoids any impossible gambles and the issues that come with them, *Theorem 1* constitutes an advance over Ramsey's system.

As noted earlier, (GRS 2) plays a very similar foundational role to (RAM 1); indeed, it is involved in every major step of the theorem's proof. It is worth noting again that, on the assumption that Ramsey wanted to avoid impossible gambles, his system requires something stronger than just (RAM 1). In fact, given his definition of $=^d$, Ramsey's system needs the stronger claim that every *outcome* is compatible with the truth or falsity of at least one ethically neutral proposition of probability ½. This can be phrased as an existential condition that replaces (RAM 1):

(RAM 1*) For every $w \in \boldsymbol{W}$, there is at least one ethically neutral proposition of probability ½, P, such that w is compatible with P and $\neg P$

(GRS 2) posits a substantially more plausible requirement than this. It postulates the existence of a set of propositions, Π, such that all have

 © Elliott 2016

probability ½, but *none* has to be ethically neutral in any of the senses defined above.[13] It is not required that every outcome has to be compatible with both the truth and the falsity of at least one proposition in Π; in fact *none* will be, if ***W*** is the set of possible worlds. Instead, it implies that for every w in ***W***, there is a proposition π in Π such that for some pair of outcomes w' and w'' with the same desirability as w, w' implies π and w'' implies $\neg\pi$. Given this, independent of its intrinsic plausibility, the use of (GRS 2) as the basis for a representation theorem constitutes a substantial advance over Ramsey's system.

(GRS 2) is nevertheless likely to be a somewhat contentious axiom. It implies, for instance, that *every* value $\underline{w}$ contains two worlds w and w' that differ with respect to some ½ probability proposition. It is plausible that for *many* values — perhaps even most — we will be able to find such a proposition. Consider, for instance, the following situation. World w is a world much like our own, where at some point in history far away a fair coin was tossed and it landed heads. No bets were ever made on the outcome of the toss, indeed nobody within the world paid attention to the outcome, and the toss had no significant impact on history in any way. World w' is essentially just like w, but the coin lands tails in it. Our subject has no intrinsic interest in the outcomes of coin tosses. Let P be the proposition *the tossed coin lands heads*. Plausibly, for our subject, $w \sim w'$, while $w \in P$ and $w' \in \neg P$. Importantly, it does not matter here that at some worlds (or even *most* worlds) the truth or falsity of P might make a difference to how our subject values that world. The propositions required by (GRS 2) might be highly relevant to the agent's subjective valuation of a world.

The case just given suggests that for most pairs of worlds w_1 and w_2, we should be able to find a ½ probability proposition which satisfies the conditions of (GRS 2). The axiom seems to be at least approximately satisfied in this sense — for any world in which there are coin tosses, we should be able to find a world which is equivalent in all respects that the agent cares about but where the outcome of some fair

[13] Suppose we define a new notion, *ethical substitutability*: P is ethically substitutable with respect to a value $\underline{w}$ iff $\underline{w}$ is compatible with P and $\neg P$, and $(P \,\&\, \underline{w}) \sim \underline{w} \sim (\neg P \,\&\, \underline{w})$. Given the (very plausible) assumption that for all values $\underline{w}$, $w \sim \underline{w}$ and if Q is a non-empty subset of $\underline{w}$, then $Q \sim \underline{w}$, (GRS 2) implies the existence of at least one ethically substitutable proposition. This is hardly problematic, however: all that is required for P to be ethically substitutable relative to some $\underline{w}$ is the existence of two equally valued worlds that differ with respect to P. Supposing that ethically substitutable propositions exist would not leave us with the problems discussed in §3.2. Thanks to Rachael Briggs and David Chalmers here.

 © Elliott 2016

coin toss is altered. If we come to a world where coin tosses matter (or where there are no coins), we should usually be able to find some other even-probability event of no intrinsic interest that we can appeal to — or some other pair of worlds with the same value which do differ with respect to some ½ probability proposition.

But it is still not obviously the case that this holds for *every* value $\underline{w}$. Perhaps there are some worlds which are unique in their desirability ranking, being equal in value to no other worlds; or perhaps there are some values which contain multiple worlds, but none of which disagree with respect to any ½ probability proposition. However, this circumstance would seem to be rare if it occurs at all, and if so it would not be a devastating problem: it would primarily mean that sometimes, $\geq^{d}$ on $\boldsymbol{W} \times \boldsymbol{W}$ is undefined. Some pairs of worlds might be left out of the $\geq^{d}$ comparison, but the relation would nevertheless still be a well-defined order on the others. It would likely be possible (though not without substantial added complexity) to prove a weaker representation result, which leaves certain utility values for worlds (and correspondingly, probability values for propositions) unspecified or within certain constrained intervals.

(GRS 3) corresponds closely to (RAM 3), and it is a standard necessary condition in decision theoretic representation theorems. The role of (GRS 4) is complex, and while no axiom like it shows up in Ramsey's system, amongst other things it plays many of the same roles as (RAM 2). Like (GRS 2), this axiom shows up in every major step of the proof that $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ is an algebraic difference structure. In a manner of speaking, this axiom says that the rational agent treats the same way all prospects with similarly valued outcomes conditional on any ½ probability proposition. It tells us that we can substitute one world w_1 for another w_1' within a gamble, or one ½ probability proposition for another, so long as the outcomes have the same desirabilities and the substitution results in a possible gamble. So, for example, if w_1 and w_1' have the same desirability and both are compatible with the ½ probability propositions π and π', then $(w_1, \pi; w_2) \sim (w_1', \pi'; w_2)$. It also allows that we can change the order of outcomes, in the sense that if $(w_1, \pi; w_2)$ and $(w_2', \pi; w_1')$ are both possible gambles, then $(w_1, \pi; w_2) \sim (w_2', \pi; w_1')$. (GRS 4) helps to ensure the coherence of the definitions of $\Pi, \geq^{d}$, and of the *bel* function.

(GRS 5) is designed to play the same role as (RAM 4): in light of the other axioms, it effectively asserts that $\geq^{d}$ is transitive, which is crucial for establishing that $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ satisfies (ADS 1) and (ADS 3) of *Definition 13*. (GRS 6) is essentially a restatement of (RAM 5). Its role is

© Elliott 2016

limited to establishing that $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ satisfies (ADS 4), a non-necessary structural condition. (GRS 7) is the Archimedean axiom; appropriately translated, it basically asserts that $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ satisfies (ADS 5).

(GRS 1)–(7) are sufficient to establish that $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ is an algebraic difference structure, which entails the existence of a real-valued function *des* on $\boldsymbol{W}$ with the property that $(w_1, w_2) \geq^{d} (w_3, w_4)$ if and only if $des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$. (GRS 8)–(9) are used to ensure that *des* represents $\succcurlyeq$ on $\boldsymbol{W} \cup \boldsymbol{G}$ in the sense that for all $\alpha, \beta \in \boldsymbol{W} \cup \boldsymbol{G}$, $des(\alpha) \geq des(\beta)$ if and only if $\alpha \succcurlyeq \beta$. These final two axioms also play central roles in the construction of a credence function *bel* (§4.4).

(GRS 8) does not correspond directly to any of Ramsey's stated axioms or any of the assumptions he otherwise mentions, though he clearly presupposed something like it in any case. It essentially states that the desirability of a gamble $(w_1, P; w_2)$ is somewhere weakly between the desirability of w_1 and the desirability of w_2. This ensures that *des* is such that for all $w_1, w_2 \in \boldsymbol{W}$, $des(w_1) \geq des(w_2)$ if and only if $w_1 \succcurlyeq w_2$. It also helps to ensure that *bel* does not result in probabilities less than 0 or greater than 1.

The sole formal role of (GRS 9) is to ensure that we can extend *des* on $\boldsymbol{W}$ to $\boldsymbol{W} \cup \boldsymbol{G}$; it is perhaps identical to what Ramsey intended for his (RAM 7). It necessitates the existence, for each gamble, of an outcome that is directly comparable with that gamble. This ensures that for all $\alpha, \beta \in \boldsymbol{W} \cup \boldsymbol{G}$, $des(\alpha) \geq des(\beta)$ if and only if $\alpha \succcurlyeq \beta$. Given the non-triviality of $\succ$ on $\boldsymbol{W} \cup \boldsymbol{G}$ (ensured by (GRS 2)) and given that, if $w_1 \succ w_2$, then $w_1 \succ (w_1, \pi; w_2) \succ w_2$, (GRS 9) forces the set of outcomes to be infinite. In this respect, it is similar to Ramsey's (RAM 6), though it plays a quite different role from that which Ramsey had intended for his axiom. This is also likely to be a contentious axiom; though here it is noteworthy that the assumption is not necessary for the main representation result. The failure of (GRS 9) would open up the possibility of *bel* being undefined for some $P \in \boldsymbol{P}$. Other means of extending *des* to $\boldsymbol{W} \cup \boldsymbol{G}$ are also likely possible in lieu of (GRS 9).

Although *Theorem 1* takes as outcomes a set of worlds, an essentially identical theorem exists where $\boldsymbol{W}$ is a set of propositions: $\boldsymbol{W} \subseteq \boldsymbol{P}$. On this interpretation, $\boldsymbol{W}$ might refer to sets of $\sim$-equivalence classes of worlds. Alternatively, $\boldsymbol{W}$ might be taken to be an arbitrary set of propositions, perhaps more or less equivalent to $\boldsymbol{P}$ itself. Formally, the difference between the theorems essentially just involves

 © Elliott 2016

substituting every instance of $\in$ between worlds and propositions to a $\subseteq$ relation (between two propositions). For instance, (GRS 2) becomes:

(GRS 2*) For every pair of propositions w_1, $w_2 \in \boldsymbol{W}$, there exists a $\pi \in \Pi$ such that for some $w_1' \sim w_1$ and $w_2' \sim w_2$, $w_1' \subseteq \pi$ and $w_2' \subseteq \neg\pi$

The representation result can then be established along precisely the same lines as in *Theorem 1, mutatis mutandis.* Of course, this would alter the interpretation of the axioms. (GRS 2*), for instance, implies that for every proposition w in $\boldsymbol{W}$, there exist at least two propositions w' and w'' and a ½ probability proposition π such that $w \sim w' \sim w''$ and $w' \subseteq \pi$ and $w'' \subseteq \neg\pi$. The modified version of the first axiom, (GRS 1), would imply that $\boldsymbol{G}$ is the set of gambles with outcomes that entail their conditions rather than merely being compatible with them.

4.4 Constructing bel

Let us suppose that $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure; our goal then is to construct a representation of the agent's credences. Closely following Ramsey's suggestion, we can define the function *bel* as follows:

Definition 14: *bel*
For all $P \in \boldsymbol{P}$, if $P = \varnothing$, then $bel(P) = 0$; and if $P = \boldsymbol{W}$, then $bel(P) = 1$; otherwise, if $w_1, w_2 \in \boldsymbol{W}$ are such that $\neg(w_1 \sim w_2)$ and $(w_1, P; w_2) \in \boldsymbol{G}$, then $bel(P) = (des((w_1, P; w_2)) - des(w_2)) / (des(w_1) - des(w_2))$

As with Ramsey's definition, *bel* so defined is unique, due to the fact that ratios of differences are preserved across admissible transformations of the *des* function.

We will also need the following condition to ensure the coherence of the definition:

Condition 1: *bel coherence*
For all $(w_1, P; w_2), (w_3, P; w_4) \in \boldsymbol{G}$ where $\neg(w_1 \sim w_2)$ and $\neg(w_3 \sim w_4)$, $(des((w_1, P; w_2)) - des(w_2)) / (des(w_1) - des(w_2)) = (des((w_3, P; w_4)) - des(w_4)) / (des(w_3) - des(w_4))$

Condition 1 is a formal restatement of one of the conditions that Ramsey briefly mentions are required to ensure the coherence of the *bel* function. We can understand what it says as follows. *Definition 14*

 © Elliott 2016

tells us that *bel*(*P*) is, say, 0.75, if it is the case that $w_1 \succ w_2$ and the value of the gamble $(w_1, P; w_2)$ sits exactly three quarters of the way from the values of w_2 to w_1. *Condition 1* then tells us that for all w_3, w_4 such that $w_3 \succ w_4$, if the gamble $(w_3, P; w_4)$ exists then it also sits three quarters of the way between w_4 and w_3 in the agent's desirability scale (and if $w_4 \succ w_3$, then $(w_3, P; w_4)$ is one quarter of the distance between w_4 and w_3). This directly implies that the value *bel*(*P*) does not depend on which outcomes and gambles we choose to consider, which is an obvious coherence requirement.[14] The other condition mentioned by Ramsey — that we will always be able to find outcomes satisfying the conditions of *Definition 14* — follows already from (GRS 1)–(3).

Theorem 3: Generalized Ramseyian credence and utility measurement If $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure and *Condition 1* holds, then there is a function *des*: $\boldsymbol{W} \cup \boldsymbol{G} \mapsto \mathbb{R}$, and a function *bel*: $\boldsymbol{P} \mapsto [0,1]$, such that for all α, β in $\boldsymbol{W} \cup \boldsymbol{G}$, all $w_1, w_2, w_3, w_4 \in \boldsymbol{W}$, and all $(w_1, P; w_2) \in \boldsymbol{G}$,

(a) $\alpha \succcurlyeq \beta$ if and only if $des(\alpha) \geq des(\beta)$

(b) $(w_1, w_2) \succeq^{d} (w_3, w_4)$ if and only if $des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$

(c) $bel(\boldsymbol{W}) = 1$ and $bel(\varnothing) = 0$

(d) $des((w_1, P; w_2)) = des(w_1).bel(P) + des(w_2).(1 - bel(P))$

Furthermore, *bel* is unique and *des* is unique up to positive linear transformation

There are some important points to note about this representation result, the proof of which is in the appendix. The first is that the function *bel* need not be a probability function, though it could be. It *is* a credence function in the sense that *bel* maps propositions to some value within [0,1]; thus it is plausible to take *bel*(*P*) as a representation of the degree of belief the agent associates with the proposition *P*.

[14] I have chosen to state *Condition 1* in terms of *des* as there is no apparent straightforward means of stating it purely in terms of preferences. Since *des* is constructed entirely from preferences, *Condition 1* is equivalent to *some* (perhaps infinitary) condition on preferences. Importantly, *Condition 1*'s content is more transparent when expressed in terms of *des*, which requires that *des* has already been characterized. Davidson and Suppes's (1956) axiom *A10* achieves a similar purpose as my *Condition 1* without referring to the intended representation, but only through a complicated series of definitions that serve to obscure its content.

 © Elliott 2016

The reason for *bel*'s permissiveness is that (GRS 1)–(9) and *Condition 1* jointly place very few restrictions on preferences for gambles conditional on propositions outside Π. For instance, suppose that neither P nor Q are in Π, $P \subset Q$, $\neg(w_1 \sim w_2)$, and the agent is to rank the two gambles $(w_1, P; w_2)$ and $(w_1, Q; w_2)$. Only (GRS 3), (GRS 8)–(9), and *Condition 1* can have any impact on how these gambles are ranked, as the other conditions are either purely existential or refer to gambles conditional on ½ probability propositions. (GRS 9) only asserts the existence of some w_3 and w_4 such that $w_3 \sim (w_1, P; w_2)$ and $w_4 \sim (w_1, Q; w_2)$, while (GRS 8) only asserts that $(w_1, P; w_2)$ and $(w_1, Q; w_2)$ must be valued somewhere between w_1 and w_2. Each of these, along with (GRS 3), can clearly be satisfied even if $(w_1, P; w_2) \succ (w_1, Q; w_2)$. Finally, *Condition 1* only restricts the relative rankings of gambles conditional on the *same* proposition, so it is also consistent with $(w_1, P; w_2) \succ (w_1, Q; w_2)$. Assuming all the other conditions to be satisfied, it follows immediately that if $(w_1, P; w_2) \succ (w_1, Q; w_2)$, then $bel(P) > bel(Q)$. Hence, *bel* in this instance is not a probability function.

It is of course possible to state further conditions to ensure that *bel* satisfies certain desirable structural properties. For instance, an extremely plausible condition is that the order in which outcomes are presented in a gamble makes no difference to their value:

Condition 2: Order indifference
For all $(w_1, P; w_2), (w_2, \neg P; w_1) \in \boldsymbol{G}$, $(w_1, P; w_2) \sim (w_2, \neg P; w_1)$

Another way to motivate *Condition 2* would be to say that $(w_1, P; w_2)$ and $(w_2, \neg P; w_1)$ are merely notational variants each representing the same object of preference; in this case, the condition will be satisfied by all preference orderings necessarily. In light of the other conditions, it implies that *bel* is such that $bel(P) = 1 - bel(\neg P)$. From (GRS 1), we know that $(w_1, P; w_2) \in \boldsymbol{G}$ if and only if $(w_2, \neg P; w_1) \in \boldsymbol{G}$; however, the axioms (GRS 1)–(9) (plus *Condition 1*) are compatible with the agent's valuing them differently (for essentially the reasons just outlined). This leads us to the following result:

Theorem 4: Generalized Ramseyian measurement with order indifference
If $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure and *Condition 1* and *2* hold, then there is a function *des*: $\boldsymbol{W} \cup \boldsymbol{G} \mapsto \mathbb{R}$, and a function *bel*: $\boldsymbol{P} \mapsto [0,1]$, such that for all α, β in $\boldsymbol{W} \cup \boldsymbol{G}$, all $w_1, w_2, w_3, w_4 \in \boldsymbol{W}$, all $(w_1, P; w_2) \in \boldsymbol{G}$, and all $P \in \boldsymbol{P}$,

(a) $\alpha \succcurlyeq \beta$ if and only if $des(\alpha) \geq des(\beta)$

 © Elliott 2016

(b) $(w_1, w_2) \geq^{d} (w_3, w_4)$ if and only if $des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$

(c) $bel(\boldsymbol{W}) = 1$ and $bel(\varnothing) = 0$

(d) $des((w_1, P; w_2)) = des(w_1).bel(P) + des(w_2).(1 - bel(P))$

(e) $bel(P) = 1 - bel(\neg P)$

Furthermore, *bel* is unique and *des* is unique up to positive linear transformation

Note that the fact that $(w_1, P; w_2)$ and $(w_2, \neg P; w_1)$ are distinct entities in $\boldsymbol{G}$ is an artefact of how we have formalized gambles, that is, as elements of $\boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{W}$. Without changing the axioms and with almost no changes to the proof, it is entirely possible to characterize $\boldsymbol{G}$ as a set of one- or two-valued functions from $\boldsymbol{W}$ to $\boldsymbol{W}$. On this formalization, $(w_1, P; w_2)$ and $(w_2, \neg P; w_1)$ will be the same object — that is, each will be a function f: $\boldsymbol{W} \mapsto \boldsymbol{W}$ such that $f(w)$ equals w_1 if $w \in P$, and w_2 if $w \in \neg P$. From this, *Condition 2* follows immediately from (GRS 3). I have chosen to use the more general formalization, $\boldsymbol{G} \subseteq \boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{W}$, as it allows for the more general representation result of *Theorem 3*. If we desire that *bel* satisfies condition (e), *Condition 2* is only a very weak additional constraint on preferences.

It would also be possible to ensure that *bel* is monotonic, in the sense that if $P \subseteq Q$, then $bel(P) \leq bel(Q)$, by adding the following condition for all $P, Q \in \boldsymbol{P}$:

Condition 3: Monotonicity
If $P \subseteq Q$, then for all $w_1, w_2 \in \boldsymbol{W}$, (i) if $w_1 \succ w_2$ and some $(w_1, Q; w_2)$, $(w_1', P; w_2')$ exists in $\boldsymbol{G}$, then $(w_1, Q; w_2) \succcurlyeq (w_1', P; w_2')$; (ii) if $w_2 \succ w_1$ and some $(w_1, Q; w_2)$, $(w_1', P; w_2')$ exists in $\boldsymbol{G}$, then $(w_1', P; w_2') \succcurlyeq (w_1, Q; w_2)$

If *Condition 3* holds and $\boldsymbol{P}$ is the power set of $\boldsymbol{W}$, then *bel* is a Choquet capacity; that is, a function f: $2^{\boldsymbol{W}} \mapsto [0,1]$ such that $f(\varnothing) = 0$, $f(\boldsymbol{W}) = 1$, and $f(Q) \geq f(P)$ whenever $P \subset Q$. (GRS 1)–(9) plus *Condition 1* and *3* thus form the basis for a version of Choquet expected utility theory (cf. Schmeidler 1989):

Theorem 5: Ramseyian Choquet expected utility theory
If $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure where $\boldsymbol{P}$ is the power set of $\boldsymbol{W}$ and *Condition 1* and *3* hold, then there is a function *des*: $\boldsymbol{W} \cup \boldsymbol{G} \mapsto \mathbb{R}$, and a function *bel*: $\boldsymbol{P} \mapsto [0,1]$, such that such that

 © Elliott 2016

for all α, β in $\boldsymbol{W} \cup \boldsymbol{G}$, all w_1, w_2, w_3, $w_4 \in \boldsymbol{W}$, all $(w_1, P; w_2) \in \boldsymbol{G}$, and all $P, Q \in \boldsymbol{P}$,

(a) $\alpha \succcurlyeq \beta$ if and only if $des(\alpha) \geq des(\beta)$

(b) $(w_1, w_2) \geq^{d} (w_3, w_4)$ if and only if $des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$

(c) $bel(\boldsymbol{W}) = 1$ and $bel(\varnothing) = 0$

(d) $des((w_1, P; w_2)) = des(w_1).bel(P) + des(w_2).(1 - bel(P))$

(f) If $P \subseteq Q$, $bel(P) \leq bel(Q)$

Furthermore, *bel* is unique and *des* is unique up to positive linear transformation

Condition 2 and *Condition 3* are independent; if both hold (and $\boldsymbol{P}$ is the power set of $\boldsymbol{W}$), then *bel* will satisfy both conditions (e) and (f).

Substantially more is required to ensure that *bel* is additive, and it is difficult to come up with conditions which are not simply direct assertions of the additivity requirement itself, or which do not involve gambles more complex than the simple two-outcome gambles we have considered so far. I am inclined to take *bel*'s potential lack of structure as a feature, not a bug. Plausibly, ordinary agents do not have probabilistically coherent degrees of belief, so any representation of credences which requires such coherence is flawed. The representation result of *Theorem 3* is compatible with an extremely wide range of credence functions—although it does require that if $P \in \Pi$, then $bel(P) = 0.5$, and there must be at least one $P \in \Pi$.[15] The result does imply that $bel(\varnothing) = 0$ and $bel(\boldsymbol{W}) = 1$, but it is noteworthy that these values for $\varnothing$ and $\boldsymbol{W}$ are entirely stipulative. Without this stipulation, *bel*'s uniqueness condition does not hold; however, *bel* on $(\boldsymbol{P} - \boldsymbol{W}) - \varnothing$ remains unique. It would be possible to set aside that stipulation and leave $bel(\varnothing)$ and $bel(\boldsymbol{W})$ undefined.

It is also important to note that the foregoing three representation results do not imply that the agent who satisfies the preference conditions can be represented as an expected utility maximizer generally. The simple reason for this is that satisfying (GRS 1)–(9) and *Condition 1*, *Condition 2*, or *Condition 3* is compatible with all kinds of preference patterns over more complex gambles. For instance, suppose that

[15] The reasoning of §2.3 essentially counts as a proof that if P is in Π, then $bel(P) = 0.5$. That there is at least one proposition in Π follows immediately from (GRS 2) and *Definition 8*.

 © Elliott 2016

$\{P_1, P_2, P_3\}$ is a partition of ***W***. Nothing about any of the above theorems implies that the desirability of the more complex gamble, $(w_1, P_1; w_2, P_2; w_3, P_3)$, must be a function of the *des* values of the outcomes and the *bel* values of the propositions. In particular, the theorem does not imply that:

$$des\big((w_1, P_1; w_2, P_2; w_3, P_3)\big) = des(w_1).bel(P_1) + des(w_2).bel(P_2) + des(w_3).bel(P_3)$$

In line with Ramsey's suggestion that we only presuppose the adequacy of the theory of subjective expected utility maximisation for the kinds of cases that we are concerned with, the representation result is compatible with high degrees of irrationality for choices more complex than those between simple gambles of the form $(w_1, P; w_2)$. Again, I take this to be a feature rather than a bug. Ordinary agents are not expected utility maximizers across the board, so it should not turn out that they can only be represented as such.

5. Conclusions

Despite its very early inception, there are several features that make Ramsey's system attractive, especially in comparison to later works. The theory of cardinal utility developed by John von Neumann and Oskar Morgenstern (1944) was in some respects a rediscovery of ideas already present in 'Truth and Probability', but its appeal to extrinsic probabilities limits its applicability, whereas Ramsey's system makes no such appeal and instead serves to characterize the agent's subjective probabilities.

Savage's (1954) theorem was also founded on Ramseyian ideas, and like Ramsey, Savage attempted to derive both a utility function and subjective probabilities purely from an agent's preferences. But Savage's system suffers from defects not present in Ramsey's system. Notably, Ramsey's system does not seem to require the assumption of *state neutrality*, which is a crucial element of Savage's system. Savage assumes that the desirabilities associated with outcomes (which are arbitrary objects or states of affairs in Savage's system) are entirely independent of the wider circumstances in which those outcomes are instantiated. If an outcome were, say, *going for a swim in the sea*, then state neutrality requires that we would value this state of affairs to a particular degree regardless of whether it is cold out, whether we have been swimming recently, or whether there have been recent reports of

 © Elliott 2016

vicious shark attacks nearby. The analogue of assuming state neutrality in a Ramseyian system would be the assumption of Naïve Expected Utility Theory, and we have seen that Ramsey does not make that mistake. Indeed, it is precisely because Ramsey recognized the falsity of this assumption that he introduced the notion of ethical neutrality.

Furthermore, given the plausible assumption that Ramsey wanted to avoid impossible gambles (§4.3), the outcomes of a gamble are always consistent with the conditions of the gamble; consequently, Ramsey's system seems to avoid the *constant acts* problem that plagues Savage's formalization. Savage posits a set of functions from an infinite set of states into a distinct set of outcomes (intended to represent possible actions an agent might take), but he crucially assumes that so-called *constant acts* exist — that there are actions we might perform which result in a particular outcome regardless of the state of the world we are in, even if that state of the world is *incompatible* with the outcome's obtaining.[16]

Of course, this is not to say that Ramsey's gambles have an unequivocally clear interpretation. (See §2.2, above.) If we construe Ramsey's measurement system as involving merely *hypothetical* judgements, as Bradley (2001) suggests, then perhaps we might read $(w_1, P; w_2)$ as just a conjunction of subjunctive conditionals:

If P were the case, then w_1, *and if* $\neg P$ *were the case, then* w_2

At least one problem arises from this interpretation, however. It seems plausible, as Jeffrey (1983) suggests, that the utility of a proposition should be a (credence) weighted average of the utilities of the worlds within it. Yet there is nothing in Ramsey's eight axioms to guarantee that agents' utilities for worlds (or near worlds) will align nicely with their utilities for the propositions that they are a part of. Of course, the same concern seems to apply to *Theorem 1*, perhaps suggesting the need for a further *utility consistency* condition if this interpretation is adopted.

Finally, on the assumption that his representation conjecture holds, another attractive feature of Ramsey's proposal is that it provides us with strong uniqueness results: if an agent satisfies Ramsey's axioms and the further stated conditions, then there is a unique credence function *bel*, and a utility function *des* unique up to positive linear transformation, which jointly represents her preferences for two-outcome

[16] See Joyce (1999, Ch. 3) for a damning discussion of the problem of constant acts in Savage's work.

 © Elliott 2016

gambles in expectational form. We might contrast this with Jeffrey's (1983) expected utility theorem, where his pair of functions, $<P, U>$ (analogous to our *bel* and *des*, respectively), is only unique up to a fractional linear transformation.

All of this is achieved, however, on the basis of the highly problematic (RAM 1) (or (RAM 1*)). For this reason, *Theorem 1* constitutes a substantial advance over Ramsey's proposed axiom system for the measurement of utilities, while *Theorem 3*, *Theorem 4*, and *Theorem 5* together show that (as Ramsey suggested but never proved) we can use such utility scales to measure degrees of belief with only minimal additional commitments. These theorems have all the attractive features just noted of Ramsey's system, without requiring the existence of ethically neutral propositions in any problematic sense. Certainly, there is no presupposition of the existence of ethically neutral propositions as Ramsey defined them (*Definition 5*), for we have made no use of logical atomism in setting out the theorems. And no proposition need non-trivially satisfy the highly questionable notion of ethical neutrality given in *Definition 6*; that is, there need not exist any contingent propositions P such that for all propositions Q that are compatible with both P and $\neg P$, $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$. Indeed, there need not even be any propositions that non-trivially satisfy the weaker *Definition 7*.

There are, naturally, further problems to deal with in the construction of Ramsey-style measurement systems. We have just seen that there are some concerns regarding the best interpretation of the elements of $\boldsymbol{G}$. It is also clear that there is still plenty of de-idealization required: the application of the theorem requires knowledge of preferences between far too many (indeed, *infinitely* many) pairs of gambles and outcomes for us ever to gain empirical access to all of them—and the longer we take to empirically discover an agent's preferences, the more likely her beliefs and desires are to change.[17] Furthermore, the outcomes that make up $\boldsymbol{W}$ are far too fine-grained to ever be fully conceptualized or described; we could never genuinely ask any person to consider whether they would prefer $(w_1, P; w_2)$ over $(w_3, P; w_4)$, if w_1, w_2, w_3, and w_4 are possible worlds. Relatedly, we have to contend with the idealizing assumption that if w_1 implies P, then $w_1 \sim (w_1 \,\&\, P)$. This requires a kind of logical omniscience on the part of our agent: a capacity to always recognize the logical equivalence of

[17] The Ramsey-style theorem of Davidson and Suppes (1956) was developed in part to deal with this problem.

 © Elliott 2016

w_1 and $(w_1 \,\&\, P)$. Finally, we will need to say more about those circumstances where the existential commitments (GRS 2) and (GRS 9) fail, and what kind of representations can be had in such cases.

It is difficult to see how we might complicate our axioms to encompass more realistic scenarios. Importantly, though, these are not issues specific to Ramsey's proposal — in fact they arise quite generally for any attempt to construct a measurement procedure on the basis of an expected utility representation theorem. We do know, however, that ethical neutrality is not a problem for this programme — we simply do not need a notion of ethical neutrality to measure degrees of belief and desire in much the way that Ramsey suggested.[18]

6. Appendix

The proof of *Theorem 1* proceeds as follows. First, we show that (GRS 1)–(7) jointly entail that $< \boldsymbol{W} \times \boldsymbol{W}, \geq^{d} >$ is an algebraic difference structure, allowing us to invoke *Theorem 2* giving us *des* on $\boldsymbol{W}$. (GRS 8) and (GRS 9) are then used to extend *des* to $\boldsymbol{W} \cup \boldsymbol{G}$, and it is shown that this provides us with an interval scale representation of $\succcurlyeq$ on $\boldsymbol{W} \cup \boldsymbol{G}$.[19]

It will be helpful to establish some lemmas first:

(Lemma A):
For every pair $w_1, w_2 \in \boldsymbol{W}$, there is a $(w_1', \pi; w_2') \in \boldsymbol{G}$

(1) Follows immediately from (GRS 1) and (GRS 2). ■

We thus know that universally quantified statements about possible gambles conditional on some ½ probability proposition are never trivially satisfied; so, for instance, where a step says 'for all $(w_1', \pi; w_4')$, $(w_2', \pi'; w_3') \in \boldsymbol{G}$, $(w_1', \pi; w_4') \succcurlyeq (w_2', \pi'; w_3')$', (Lemma A) ensures that at least one such pair of gambles exists in $\boldsymbol{G}$. I will generally omit this step in what follows. Set memberships have been suppressed where obvious: henceforth we are only concerned with gambles in $\boldsymbol{G}$.

[18] Thanks to Rachael Briggs, David Chalmers, Alan Hájek, Hanti Lin, Wolfgang Schwarz, a PhilSoc audience at the ANU, a 2014 AAP conference audience in Canberra, and two anonymous referees at *Mind* for helpful feedback and discussion.

[19] Several of the steps in what follows owe much to Bradley (2001), especially (Lemma C) and the steps involving it.

 © Elliott 2016

(Lemma B):
If $(w_1', \pi; w_4') \succcurlyeq (w_2', \pi'; w_3')$ for some pair $(w_1', \pi; w_4')$, $(w_2', \pi'; w_3')$, then $(w_1, w_2) \succeq^{d} (w_3, w_4)$

(1) Suppose that $(w_1', \pi; w_4') \succcurlyeq (w_2', \pi'; w_3')$ for some such pair.

(2) By (Lemma A), some $(w_4'', \pi''; w_1'')$ exists, and by successive iterations of (GRS 4), $(w_1', \pi; w_4') \sim (w_4'', \pi''; w_1'')$ and $(w_4'', \pi''; w_1'') \sim (w_1''', \pi'''; w_4''')$ for all such pairs. Because $\sim$ is an equivalence relation (GRS 3), $(w_1, \pi; w_4) \sim (w_1''', \pi'''; w_4''')$ for all such pairs.

(3) By the same steps, we know that $(w_2', \pi'; w_3') \sim (w_2'', \pi^*; w_3'')$ for all such pairs.

(4) So given our starting supposition, $(w_1''', \pi'''; w_4''') \succcurlyeq (w_2'', \pi^*; w_3'')$ for all such pairs, which is just the right hand side of *Definition 9*. ■

(Lemma C):
If $(w_1, w_2) \succeq^{d} (w_3, w_4)$, then $(w_4, w_3) \succeq^{d} (w_2, w_1)$, and $(w_1, w_3) \succeq^{d} (w_2, w_4)$

(1) Suppose $(w_1, w_2) \succeq^{d} (w_3, w_4)$, so $(w_1', \pi; w_4') \succcurlyeq (w_2', \pi'; w_3')$ for all such gambles.

(2) (Lemma A) ensures some $(w_4'', \pi^*; w_1'')$, $(w_3'', \pi^+; w_2'')$ exist, and by (GRS 4), $(w_4'', \pi^*; w_1'') \sim (w_1', \pi; w_4')$ and $(w_3'', \pi^+; w_2'') \sim (w_2', \pi'; w_3')$. Substituting for equally valued gambles, we get $(w_4'', \pi^*; w_1'') \succcurlyeq (w_3'', \pi^+; w_2'')$, which given (Lemma B) implies $(w_4, w_3) \succeq^{d} (w_2, w_1)$.

(3) Likewise, $(w_1', \pi; w_4') \succcurlyeq (w_3'', \pi^+; w_2'')$, so $(w_1, w_3) \succeq^{d} (w_2, w_4)$. ■

We can now show that (ADS 1)–(5) follow from (GRS 1)–(7). (ADS 2) is simply the first part of (Lemma C). Next we will prove that $\succeq^{d}$ on $\boldsymbol{W} \times \boldsymbol{W}$ is complete:

(1) From (Lemma A), for any two (w_1, w_4), (w_2, w_3), there exist $(w_1', \pi; w_4')$, $(w_2', \pi'; w_3')$. From (GRS 3), either $(w_1', \pi; w_4') \succcurlyeq (w_2', \pi'; w_3')$ or $(w_2', \pi'; w_3') \succcurlyeq (w_1', \pi; w_4')$.

(2) Given (Lemma B), if the former then $(w_1, w_2) \succeq^{d} (w_3, w_4)$, and if the latter then $(w_3, w_4) \succeq^{d} (w_1, w_2)$. So either $(w_1, w_2) \succeq^{d} (w_3, w_4)$ or $(w_3, w_4) \succeq^{d} (w_1, w_2)$. ■

© Elliott 2016

We also prove that $\geq^{d}$ on $\boldsymbol{W} \times \boldsymbol{W}$ is transitive:

(1) Suppose that $(w_1, w_2) \geq^{d} (w_3, w_4)$ and $(w_3, w_4) \geq^{d} (w_5, w_6)$.

(2) From *Definition* 9, for all the relevant gambles, this implies that $(w'_1, \pi; w'_4) \succcurlyeq (w'_2, \pi'; w'_3)$ and $(w'_3, \pi^{*}; w'_6) \succcurlyeq (w'_4, \pi^{+}; w'_5)$.

(3) For any pair of gambles $(w''_1, \pi''; w''_6)$, $(w''_2, \pi'''; w''_5)$, (GRS 5) then requires that $(w''_1, \pi''; w''_6) \succcurlyeq (w''_2, \pi'''; w''_5)$, and $(w_1, w_2) \geq^{d} (w_5, w_6)$ follows from (Lemma B). ■

So $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ is a weak order and (ADS 1) is satisfied. Next we show that (ADS 3) is satisfied:

(1) Suppose $(w_1, w_2) \geq^{d} (w_4, w_5)$ and $(w_2, w_3) \geq^{d} (w_5, w_6)$.

(2) The second part of (Lemma C) applied to each conjunct entails $(w_1, w_4) \geq^{d} (w_2, w_5)$ and $(w_2, w_5) \geq^{d} (w_3, w_6)$. Because $\geq^{d}$ is transitive, $(w_1, w_4) \geq^{d} (w_3, w_6)$. So from (Lemma C) again, $(w_1, w_3) \geq^{d} (w_4, w_6)$. ■

(ADS 4) is satisfied:

(1) Suppose $(w_1, w_2) \geq^{d} (w_3, w_4) \geq^{d} (w_1, w_1)$.

(2) From (GRS 6), for every triple w_1, w_3, w_4, there is a w_5 such that for some $(w'_1, \pi; w'_4)$, $(w_5, \pi'; w'_3)$ (ensured by (Lemma A)), $(w'_1, \pi; w'_4) \sim (w_5, \pi'; w'_3)$. Applying (Lemma B), we see that there must be a w_5 such that $(w_1, w_5) =^{d} (w_3, w_4)$.

(3) Likewise, for every triple w_3, w_4, w_2, there is a w_6 such that $(w'_3, \pi; w'_2) \sim (w_6, \pi'; w'_4)$ for some such pair; so there is a w_6 such that $(w_3, w_4) =^{d} (w_6, w_2)$. ■

And (ADS 5) is also satisfied. The proof of this is trivial given (GRS 7), and *Definition* 9, and the definition of a strictly bounded standard sequence; it has therefore been left unstated. (GRS 1)–(7) therefore imply that $<\boldsymbol{W} \times \boldsymbol{W}, \geq^{d}>$ is an algebraic difference structure, which ensures the existence of *des* on $\boldsymbol{W}$ (unique up to positive linear transformation), such that:

$$(w_1, w_2) \geq^{d} (w_3, w_4) \text{ if and only if } des(w_1) - des(w_2) \geq des(w_3) - des(w_4)$$

 © Elliott 2016

We appeal primarily to (GRS 8) to show that $des(w_1) \geq des(w_2)$ if and only if $w_1 \succcurlyeq w_2$:

(1) From (GRS 8), $w_1' \sim w_1$ if and only if, for all $(w_1'', P; w_1''')$, $w_1 \sim (w_1'', P; w_1''')$; and similarly, $w_2' \sim w_2$ if and only if, for all $(w_2'', P; w_2''')$, $w_2 \sim (w_2'', P; w_2''')$.

(2) Given (GRS 3) then, $w_1 \succcurlyeq w_2$ if and only if $(w_1'', \pi; w_1''') \succcurlyeq (w_2'', \pi'; w_2''')$ for all such gambles, which holds if and only if $(w_1, w_2) \geq^{d} (w_2, w_1)$.

(3) From *Theorem 2*, $(w_1, w_2) \geq^{d} (w_2, w_1)$ if and only if $des(w_1) - des(w_2) \geq des(w_2) - des(w_1)$, which can only be if $des(w_1) \geq des(w_2)$. So $w_1 \succcurlyeq w_2$ if and only if $des(w_1) \geq des(w_2)$. ■

We further require that *des* is defined on $\boldsymbol{W} \cup \boldsymbol{G}$. From (GRS 9), we know that for every $(w_1, P; w_2)$ there is a w_3 such that $(w_1, P; w_2) \sim w_3$. We can achieve the desired extension by making the following stipulation:

For all w_3, $(w_1, P; w_2) \in \boldsymbol{W} \cup \boldsymbol{G}$, $des((w_1, P; w_2)) = des(w_3)$ if and only if $(w_1, P; w_2) \sim w_3$

The proof that condition (a) of *Theorem 1* then holds is trivial and left unstated. The uniqueness properties of *des* on $\boldsymbol{W}$ will also clearly hold for *des* on $\boldsymbol{W} \cup \boldsymbol{G}$. The foregoing thus establishes *Theorem 1*.

To prove *Theorem 3*, we need to show that *Definition 14* provides us with a unique function *bel* that satisfies the stated properties. There are three parts to proving this. First, we need to show that we will always be able to find worlds and gambles satisfying the definition's conditions. Second, we need to show that the $bel(P)$ is independent of the choice of worlds and gambles satisfying the definition; along with the first step this will ensure that *bel* is a function with domain $\boldsymbol{P}$. Finally, we show that the range of the function is [0,1].

(1) By stipulation, if $P = \varnothing$ then $bel(P) = 0$; and if $P = \boldsymbol{W}$, then $bel(\boldsymbol{W}) = 1$. Thus we are only concerned with $bel(P)$ for contingent propositions, which must contain at least one member.

(2) Given (GRS 2) and *Definition 8*, $\succ$ on $\boldsymbol{W}$ is non-trivial. Let w_1 and w_2 be two outcomes such that $\neg(w_1 \sim w_2)$. There are then three possibilities: for any contingent proposition P, either P

 © Elliott 2016

contains both w_1 and w_2, or $\neg P$ does, or P and $\neg P$ contain one of w_1 and w_2 each.

(3) If either P contains both outcomes, given (GRS 3), its negation must contain at least one world w_3 such that either $\neg(w_1 \sim w_3)$ or $\neg(w_2 \sim w_3)$. Likewise if $\neg P$ contains both outcomes. If P and $\neg P$ contain one of w_1 and w_2 each, we already know that there are worlds in P and in $\neg P$ such that the agent is not indifferent between them. In any case, then, for any contingent proposition P, given (GRS 1), there will be at least one gamble of the form $(w_1, P; w_2)$ where $\neg(w_1 \sim w_2)$.

(4) That $bel(P)$ is independent of the choice of worlds and gambles satisfying the antecedent conditions follows immediately from *Condition 1*.

(5) The range of *bel* is [0,1]: from (GRS 8) and (GRS 3), for all $(w_1, P; w_2)$, either $w_1 \succcurlyeq w_2$ and $w_1 \succcurlyeq (w_1, P; w_2) \succcurlyeq w_2$, or $w_2 \succcurlyeq w_1$ and $w_2 \succcurlyeq (w_1, P; w_2) \succcurlyeq w_1$. Given the established properties of *des*, we know $des(w_1, P; w_2)$ always sits somewhere weakly between $des(w_1)$ and $des(w_2)$. It follows that the ratio of the difference between $des((w_1, P; w_2))$ and $des(w_2)$ and the difference between $des(w_1)$ and $des(w_2)$ will always be within [0,1]. ■

We now prove that *bel* is such that $des((w_1, P; w_2)) = des(w_1).bel(P) + des(w_2).(1 - bel(P))$:

(1) Suppose first that $w_1 \sim w_2$; then, by reasoning noted above, $des(w_1) = des(w_2) = des((w_1, P; w_2))$. Let $des(w_1) = x$. The required equality then holds just in case $x = x.bel(P) + x.(1 - bel(P))$; we have already noted that $bel(P) \in [0,1]$, so this is true regardless of the value of $bel(P)$.

(2) Suppose next that $\neg(w_1 \sim w_2)$. From *Definition 14*, $bel(P) = (des((w_1, P; w_2)) - des(w_2)) / (des(w_1) - des(w_2))$, which holds if and only if

$(des(w_1) - des(w_2)).bel(P) = des((w_1, P; w_2)) - des(w_2)$ if and only if

$des((w_1, P; w_2)) = des(w_1).bel(P) - des(w_2).bel(P) + des(w_2)$ if and only if

 © Elliott 2016

$$des((w_1,\ P;\ w_2)) = des(w_1).bel(P) + des(w_2).(1 - bel(P)).\ \blacksquare$$

Finally, *bel* is unique:

(1) From the earlier proofs and the fact that ratios of differences are preserved across admissible transformations of *des*, we know that there is only one function *bel* on $\boldsymbol{P}$ such that $bel(\boldsymbol{W}) = 1$, $bel(\varnothing) = 0$, and for any contingent P, if w_1, w_2 are such that $\neg(w_1 \sim w_2)$ and $(w_1,\ P;\ w_2)$, then $bel(P) = des((w_1,\ P;\ w_2)) - des(w_2)\ /\ des(w_1) - des(w_2)$.

(2) We have also already established that the previous equality holds if and only if $des((w_1,\ P;\ w_2)) = des(w_1).bel(P) - des(w_2).bel(P) + des(w_2)$. Since there is only one function satisfying the left-hand side, only one satisfies the right-hand side. ■

The foregoing establishes *Theorem 3*. Next we prove *Theorem 4*. Clearly, assuming that $<\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{G}, \succcurlyeq>$ is a generalized Ramsey structure and *Condition 1* holds, the only thing we need to prove here is that if *Condition 2* holds then $bel(P) = 1 - bel(\neg P)$.[20]

(1) $bel(\boldsymbol{W}) = 1$ and $bel(\varnothing) = 0$ by stipulation, so we are only interested in contingent P.

(2) As already shown, for all contingent P, there is some $(w_1,\ P;\ w_2)$ such that the ratio $(des((w_1,\ P;\ w_2)) - des(w_2))\ /\ (des(w_1) - des(w_2))$ is defined (i.e., such that $\neg(w_1 \sim w_2)$). From *Theorem 3*, this is the value of $bel(P)$.

(3) From (GRS 1) and since $\boldsymbol{P}$ is closed under negation, if $(w_1,\ P;\ w_2)$ is in $\boldsymbol{G}$ then $(w_2,\ \neg P;\ w_1)$ is in $\boldsymbol{G}$; thus $bel(\neg P) = (des((w_2,\ \neg P;\ w_1)) - des(w_1))\ /\ (des(w_2) - des(w_1))$. Multiplying the denominator and the numerator by -1 gets us $bel(\neg P) = (des(w_1) - des((w_2,\ \neg P;\ w_1)))\ /\ (des(w_1) - des(w_2))$.

(4) *Condition 2* ensures $(w_1,\ P;\ w_2) \sim (w_2,\ \neg P;\ w_1)$, so $des((w_1,\ P;\ w_2)) = des((w_2,\ \neg P;\ w_1))$.

(5) Let $des((w_1,\ P;\ w_2)) = x$. Given the foregoing, $bel(P) + bel(\neg P)$ is equal to:

$$((x - des(w_2))\ /\ (des(w_1) - des(w_2))) + ((des(w_1) - x)\ /\ (des(w_1) - des(w_2)))$$

[20] Thanks to Rachael Briggs for the main outline of the following proof.

 © Elliott 2016

$= (x - w_2 + w_1 - x) \;/\; (w_1 - w_2) = (x - x + w_1 - w_2) \;/\; (w_1 - w_2)$

$= (w_1 - w_2) \;/\; (w_1 - w_2) = 1$. Condition (e) of *Theorem 4* follows immediately. ■

Finally, we can prove *Theorem 5*. As before, given *Theorem 3*, the only thing we need to prove here is that if *Condition 3* holds along with the other conditions then *bel* is monotonic.

(1) If $P = \varnothing$, then by stipulation, $bel(P) = 0$ and for all Q, $bel(P) \leq bel(Q)$. If $Q = \boldsymbol{W}$, then $bel(Q) = 1$ and for all P, $bel(P) \leq bel(Q)$. Thus suppose that P and Q are contingent and $P \subseteq Q$.

(2) From *Condition 3*, for all w_1, w_2, if some $(w_1, Q; w_2)$, $(w_1', P; w_2')$ exist, then: if $w_1 \succ w_2$, $(w_1, Q; w_2) \succcurlyeq (w_1', P; w_2')$, and if $w_2 \succ w_1$, $(w_1', P; w_2') \succcurlyeq (w_1, Q; w_2)$.

(3) We know from steps proven above that for all contingent Q, we will find at least one pair w_1, w_2 such that either (a) $w_1 \succ w_2$ and $(w_1, Q; w_2)$ exists, or (b) $w_2 \succ w_1$ and $(w_1, Q; w_2)$ exists. Given either (a) or (b), there are only two possibilities: $(w_1', P; w_2')$ exists in $\boldsymbol{G}$, or it does not.

(4) Suppose (a) is true, and that $(w_1', P; w_2')$ exists; thus $(w_1, Q; w_2) \succcurlyeq (w_1', P; w_2')$ and therefore $des((w_1, Q; w_2)) \geq des((w_1', P; w_2'))$ and $des(w_1) > des(w_2)$. With the established properties of *bel*, the former (weak) inequality implies $des(w_1).bel(Q) + des(w_2).(1 - bel(Q)) \geq des(w_1).bel(P) + des(w_2).(1 - bel(P))$. This can only hold given the latter (strict) inequality if $bel(Q) \geq bel(P)$.

(5) Suppose now that (b) is true, and that $(w_1', P; w_2')$ exists. By the same steps just given, $des(w_1).bel(P) + des(w_2).(1 - bel(P)) \geq des(w_1).bel(Q) + des(w_2).(1 - bel(Q))$ and $des(w_1) > des(w_2)$, which can only hold if $bel(Q) \geq bel(P)$.

(6) So given either (a) or (b) and the assumption that $(w_1', P; w_2')$ exists, $bel(Q) \geq bel(P)$.

(7) Assume now instead either (a) or (b), and that $(w_1', P; w_2')$ is not in $\boldsymbol{G}$. Because $(w_1, Q; w_2)$ is in $\boldsymbol{G}$ and all members of $\neg Q$ are members of $\neg P$, we know that there is some $w_2' \in \neg P$. Thus $(w_1', P; w_2') \notin \boldsymbol{G}$ only if there is no $w_1' \in P$. Since all

 © Elliott 2016

members of P are members of Q, Q must contain members of multiple values.

(8) All $w \in P$ must be such that for all $w' \in \neg Q$, $w \sim w'$; if instead $\neg(w \sim w')$, then since any member of P is also in Q and any member of $\neg Q$ is also in $\neg P$, we would have $w \succ w'$ and $(w, P; w')$, $(w, Q; w')$ in $\boldsymbol{G}$ or $w' \succ w$ and $(w, P; w')$, $(w, Q; w')$ in $\boldsymbol{G}$, contradicting the earlier supposition. Given this, Q must contain all $w \in \boldsymbol{W}$ such that $\neg(w \sim w')$, for any $w' \in P$.

(9) Given (GRS 2), all $\sim$-equivalence classes of worlds contain at least two members; thus, for any $w_1 \in Q$ and $w_2 \in P$ such that $w_1 \neq w_2$ and furthermore $\neg(w_1 \sim w_2)$, there is a $w_1' \in Q$ which is not in P. Let R be the proposition which contains all members of P, and w_1' but not w_1. Thus, $P \subset R \subset Q$. As $\boldsymbol{P}$ is the power set of $\boldsymbol{W}$, $R \in \boldsymbol{P}$.

(10) We now know that for some w_1, w_2', $w_1 \succ w_2'$ and $(w_1, Q; w_2')$, $(w_1', R; w_2')$ exist, or $w_2' \succ w_1$ and $(w_1, Q; w_2')$, $(w_1', R; w_2')$ exist. By the steps given earlier, $bel(Q) \geq bel(R)$. We also know that for some w_1, w_2, $w_1 \succ w_2$ and $(w_2, R; w_1)$, $(w_2, P; w_1)$ exist, or $w_2 \succ w_1$ and $(w_2, R; w_1)$, $(w_2, P; w_1)$ exist. By the same steps, then, $bel(R) \geq bel(P)$. By transitivity, $bel(Q) \geq bel(P)$. ■

References

Bartha, Paul 2007: 'Taking Stock of Infinite Value: Pascal's Wager and Relative Utilities'. *Synthese*, 54, pp. 5–52.

Bradley, Richard 2001: 'Ramsey and the Measurement of Belief'. In Corfield and Williamson 2001, pp. 263–90.

Chalmers, David 2011: 'Frege's Puzzle and the Objects of Credence'. *Mind*, 120, pp. 587–635.

Corfield, David and Jon Williamson (eds) 2001: *Foundations of Bayesianism*. Dordrecht: Kluwer.

Davidson, Donald, Sidney Siegel, and Patrick Suppes 1957: *Decision Making: An Experimental Approach*. Stanford, CA: Stanford University Press.

Davidson, Donald and Patrick Suppes 1956: 'A Finitistic Axiomatization of Subjective Probability and Utility'. *Econometrica*, 24, pp. 264–75.

 © Elliott 2016

Debreu, Gerard 1959: 'Cardinal Utility for Even-chance Mixtures of Pairs of Sure Prospects'. *Review of Economic Studies*, 26, pp. 174–7.

de Finetti, Bruno 1931: 'Sul Significato Soggettivo Della Probabilita'. *Fundamenta Mathematicae*, 17, pp. 298–329.

Fishburne, Peter C. 1967: 'Preference-Based Definitions of Subjective Probability'. *Annals of Mathematical Statistics*, 38, pp. 1,605–17.

Jeffrey, Richard 1983: *The Logic of Decision*, second edition. Chicago: University of Chicago Press, rpnt 1990.

Joyce, James 1999: *The Foundations of Causal Decision Theory*. New York: Cambridge University Press.

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky 1971: *Foundations of Measurement Volume 1*. New York: Academic Press.

Ramsey, Frank P. 1931a: *The Foundations of Mathematics and other Logical Essays*. New York: Harcourt, Brace and Company.

Ramsey, Frank P. 1931b: 'Truth and Probability'. In his 1931a, pp. 156–98.

Savage, Leonard J. 1954: *The Foundations of Statistics*. New York: Dover.

Schmeidler, David 1989: 'Subjective Probability and Expected Utility without Additivity'. *Econometrica*, 57, pp. 571–87.

Sobel, Jordan Howard 1998: 'Ramsey's Foundations Extended to Desirabilities'. *Theory and Decision*, 44, pp. 231–78.

Suppes, Patrick 1956: 'The Role of Subjective Probability and Utility in Decision Making'. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5, pp. 61–73.

von Neumann, John and Oskar Morgenstern 1944: *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

© Elliott 2016

Hume's Doxastic Involuntarism

Hsueh Qu
National University of Singapore
hsueh.qu@gmail.com

In this paper, I examine three mutually inconsistent claims that are commonly attributed to Hume: (a) all beliefs are involuntary; (b) some beliefs are subject to normative appraisal; and (c) that 'Ought implies Can'. I examine the textual support for such ascription, and the options for dealing with the puzzle posed by their inconsistency. In what follows I will put forward some evidence that Hume maintains each of the three positions outlined above. I then examine what I call the 'prior voluntary action' solution. I argue that this position in any form fails to account for synchronic rationality. I then raise more specific objections depending on how we disambiguate the position, which can be read as either granting beliefs derivative voluntariness, or as denying their normative significance; the former version is inconsistent with Hume's treatment of natural abilities, while the latter falls to a regress given Hume's thesis regarding the inability of actions and passions to be subject to epistemic normativity. I then propose to reject (c) instead for two reasons: first, the weakness of textual support for such an ascription; secondly, Hume's explicit recognition of the irrelevance of involuntariness to normative evaluation in the moral case.

1. Three inconsistent views

Hume is famous for providing a naturalistic account of belief-formation according to which belief is a purely involuntary phenomenon. Yet he is also famous for his disapproval of irrational beliefs held by the vulgar and the superstitious, among others. However, if we take these two positions together with the intuitive principle that 'Ought implies Can', we have an inconsistency. The contrapositive of 'Ought implies Can' is 'Cannot implies not-Ought'. Since beliefs are involuntary, we cannot believe otherwise than we do. Thus, according to the principle that 'Cannot implies not-Ought', it is not the case that we ought to hold any set of beliefs other than the set that we actually hold (that is, it is not the case that we ought to drop beliefs that we have, or hold beliefs that we do not). Therefore, in claiming that there are certain beliefs that some people ought not to hold (and therefore ought to drop), Hume is violating the contrapositive of 'Ought implies Can'.

In what follows I will put forward some evidence that Hume maintains each of the three positions outlined above. I then examine what I

Mind, Vol. 126 . 501 . January 2017
doi:10.1093/mind/fzv181 Advance Access publication 25 September 2016

© Qu 2016

call the 'Prior Voluntary Action' solution endorsed by Passmore (1980), Norton (1982, 1994, 2002), Falkenstein (1997), Owen (1999), Williams (2004), and McCormick (2005), among others. I argue that 'Prior Voluntary Action' in any form fails to account for synchronic rationality. I then raise more specific objections depending on how we disambiguate 'Prior Voluntary Action'. 'Prior Voluntary Action' can be read as either granting beliefs derivative voluntariness, or derivative normative significance; the former reading is textually unsubstantiated and indeed textually problematic, while the latter reading falls to a regress given Hume's thesis regarding the inability of actions and passions to intrinsically bear epistemic normativity.

In the end, I propose two reasons for dropping the assumption that Hume subscribes to 'Ought implies Can' in the epistemic realm: first, the weakness of textual support for Hume's subscribing to 'Ought implies Can'; secondly, Hume's recognition of the irrelevance of involuntariness to the moral evaluation of qualities of the mind.

1.1 Doxastic involuntarism

In arguing that what distinguishes an idea from a belief cannot be an additional idea annexed to the former, Hume endorses doxastic involuntarism (that is, the thesis that beliefs are involuntary):

> ... The mind has the command over all its ideas, and can separate, unite, mix, and vary them, as it pleases; so that if belief consisted merely in a new idea, annex'd to the conception, it wou'd be in a man's power to believe what he pleas'd. We may, therefore, conclude, that belief consists merely in a certain feeling or sentiment, in something, that depends not on the will, but must arise from certain determinate causes and principles, of which we are not masters. (THN App 2; SBN 623-4)[1]

> The mind has a faculty of joining all ideas together, which involve not a contradiction; and therefore if belief consisted in some idea, which we add to the simple conception, it would be in a man's power, by adding this idea to it, to believe any thing, which he can conceive.

> Since therefore belief implies a conception, and yet is something more; and since it adds no new idea to the conception; it follows, that it is a different manner of conceiving an object; *something* that is distinguishable

[1] In references to Hume's texts throughout the paper, 'THN' refers to the *Treatise of Human Nature*, 'EHU' to the *Enquiry Concerning Human Understanding*, and 'EPM' to the *Enquiry Concerning the Principles of Morals*. Arabic numerals refer to section and paragraph numbers (EHU and EPM), or to book, part, section, and paragraph numbers (THN). SBN numbers refer to pages in the Selby-Bigge and Nidditch editions of the *Treatise* and two *Enquiries*.

 © Qu 2016

to the feeling, and depends not upon our will, as all our ideas do. (*Abstract* 20-21; SBN 653, emphasis in original)

For as the mind has authority over all its ideas, it could voluntarily annex this particular idea to any fiction, and consequently be able to believe whatever it pleases; contrary to what we find by daily experience. We can, in our conception, join the head of a man to the body of a horse; but it is not in our power to believe, that such an animal has ever really existed.

It follows, therefore, that the difference between *fiction* and *belief* lies in some sentiment or feeling, which is annexed to the latter, not to the former, and which depends not on the will, nor can be commanded at pleasure. (EHU 5.10-11; SBN 48, emphasis in original)

One question is whether Hume simply means the weaker thesis that is the denial that we have a power of 'libertarian-can' with respect to beliefs—in other words, the thesis that beliefs are beyond our libertarian control (i.e. we do not possess a contra-causal power of agent causation that would allow us to contravene the physically determined outcome in the doxastic realm); or whether he means the stronger thesis that is the denial that we have a power of 'voluntary-can' with respect to our beliefs—in other words, the thesis that beliefs are beyond our voluntary control (i.e. doxastic involuntarism). I think the passages clearly state the latter. To see this, it would help to specify what involuntariness consists in for Hume. In his discussion of free will, Hume notes that the only meaningful sense of liberty is that sense that applies to voluntary actions, which is determination by the will:

For what is meant by liberty, when applied to voluntary actions? ... By liberty, then, we can only mean *a power of acting or not acting, according to the determinations of the will*; that is, if we chuse to remain at rest, we may; if we chuse to move, we also may. (EHU 8.23; SBN 95, emphasis in original)

To say that X is under my voluntary control is to say that X is suitably determined by the will.[2] Conversely, Hume characterizes involuntariness as being 'independent from the will' (EPM App 4.20) and 'beyond the dominion of the will' (EPM App 4.21). Crucially, if beliefs are not subject to determination by the will, then they are not subject to voluntary control.

[2] Work will need to be done to adequately flesh out this sense of 'suitably'; for instance, there will need to be a clause ruling out determination by deviant causal chains. Nevertheless, I gloss over the matter here. The only result I require for my purposes is that if something is not determined by the will, it is involuntary; this holds true regardless of how we flesh out the suitability clause.

 © Qu 2016

In THN App 2, Hume states that 'belief consists merely in a certain feeling or sentiment, in something, that depends not on the will, but must arise from certain determinate causes and principles, of which we are not masters'. In Abstract 21, Hume states that 'belief... is a different manner of conceiving an object; *something* that is distinguishable to the feeling, and depends not upon our will, as all our ideas do' (emphasis in original). In EHU 5.11, Hume says: 'It follows, therefore, that the difference between *fiction* and *belief* lies in some sentiment or feeling, which is annexed to the latter, not to the former, and which depends not on the will, nor can be commanded at pleasure' (emphasis in original). All three discussions of belief above emphasize that the sentiment constitutive of belief is independent of the will; in short, belief is involuntary. It is important to note that in these three passages, Hume rules out not only our wilfully believing, but also our wilfully disbelieving. Hume's central point is that 'belief consists merely in a certain feeling or sentiment... that depends not on the will' (THN App 2); since we cannot control when this sentiment is present *or absent*, we can neither control our believing nor our disbelieving.

The point can also be established by examining Hume's argument in all three passages. Hume's point in all three passages is that belief cannot consist in an additional idea annexed to the believed idea, because this would entail the false thesis that we could believe whatever we like. This is because, Hume says, 'the mind has the command over all its ideas, and can separate, unite, mix, and vary them, as it pleases' (THN App 2); 'the mind has authority over all its ideas' (EHU 5.11); 'The mind has a faculty of joining all ideas together, which involve not a contradiction' (*Abstract* 20). Clearly, Hume means that the mind has *voluntary* control over its ideas; he certainly does not mean that the mind has *libertarian* control over its ideas, because the mind does not have libertarian control over *anything*, since it is determined; recall Hume's words in EHU 8.23 cited earlier. If this is the case, Hume must mean that belief is beyond our *voluntary* control. The sense of control Hume appeals to in both premises must be the same one; otherwise, the conclusion that belief cannot be an additional idea annexed to the believed idea would not follow. Indeed, in these three passages, Hume finds doxastic involuntarism so obvious that he does not even bother arguing for it at this point in his dialectic—as Passmore (1980, p. 160) puts it, 'the *reductio* is complete' for Hume, who takes it to be patently clear that the distinction between beliefs and ideas is not grounded on their contents, given the obviousness of doxastic involuntarism.

 © Qu 2016

Moreover, there is further textual evidence that belief is involuntary for Hume. In THN 3.1.1.12, Hume argues that mistaken beliefs are not morally culpable, pointing out that moralists do not typically consider them criminal because they are 'perfectly involuntary'. Moreover, Hume compares beliefs to the passions with regard to their involuntariness: '… belief … is an operation of the soul, when we are so situated, as unavoidable as to feel the passion of love, when we receive benefits; or hatred, when we meet with injuries' (EHU 5.8). All in all, there seems extremely strong textual evidence that Hume thinks belief is involuntary.

It is worth further investigating the nature of Hume's doxastic involuntarism. As stated above, beliefs are involuntary for Hume because they are characterized by 'some sentiment or feeling', which 'depends not on the will' (EHU 5.11). Our beliefs are involuntary, because we cannot generate this unique sentiment (i.e. force and vivacity) merely by willing to. This is not to say that we cannot, through more indirect means, manipulate the circumstances under which this sentiment arises. Crucially, reasoning consists in the manipulation of ideas in such a way as to generate beliefs, for instance. 'The mind has authority over all its ideas' (EHU 5.10), and so the act of putting ideas together is a free and voluntary one. In his discussion of miracles, Hume memorably describes the reasoning involved in what amounts to crude enumerative induction as consisting in arranging certain relevant ideas:

> A wise man, therefore, proportions his belief to the evidence. In such conclusions as are founded on an infallible experience, he expects the event with the last degree of assurance, and regards his past experience as a full *proof* of the future existence of that event. In other cases, he proceeds with more caution: He weighs the opposite experiments: He considers which side is supported by the greater number of experiments: To that side he inclines, with doubt and hesitation; and when at last he fixes his judgment, the evidence exceeds not what we properly call *probability*. All probability, then, supposes an opposition of experiments and observations, where the one side is found to overbalance the other, and to produce a degree of evidence, proportioned to the superiority. A hundred instances or experiments on one side, and fifty on another, afford a doubtful expectation of any event; though a hundred uniform experiments, with only one that is contradictory, reasonably beget a pretty strong degree of assurance. In all cases, we must balance the opposite experiments, where they are opposite, and deduct the smaller number from the greater, in order to know the exact force of the superior evidence. (EHU 10.4; SBN 112, emphasis in original)

 © Qu 2016

Here Hume describes the probable reasoning of a wise man as putting the ideas of similar outcomes together and contrasting these with the ideas of differing outcomes. Of course, note that these ideas of outcomes have themselves to be vivacious in order to carry weight in this process. This arrangement of ideas is a voluntary act; however, the output—that is, the sentiment of belief—is nevertheless involuntary. For instance, shortly after the above passage Hume states: 'Where this experience is not entirely uniform on any side, it is attended with an *unavoidable* contrariety in our judgments' (EHU 10.6, emphasis added).

Just as one can manipulate the circumstances under which we believe something, we can also manipulate the circumstances under which we doubt something. Doubt itself, being the diminishing of vivacity, is an involuntary phenomenon. However, by voluntarily considering countervailing evidence or the sorry state of our faculties, we can nevertheless attempt to diminish the vivacity of a certain idea. An example of this would be Hume's mitigated scepticism, which suggests that we take 'a small tincture of Pyrrhonism' to induce 'a degree of doubt, and caution, and modesty' (EHU 12.24). Clearly, there are many voluntary mental acts that lead to both belief and doubt, and these form an important part of Hume's discussions of reasoning and scepticism.

Despite our ability to manipulate the sentiment that constitutes belief, it should be emphasized that this sentiment is nevertheless wholly involuntary. It is helpful to compare vivacity with other sentiments. I may be able to manipulate my stimuli through various physical and mental acts in order to engender certain passions: I know that if I visit a website espousing racism or misogyny, I will experience anger; if I recollect the scene of Fantine's death, I will feel sadness; if I stand on the rooftop of a high building, I will feel fear. This limited ability to manipulate our sentiments through changing our stimuli does not change the fact that these sentiments are wholly involuntary. Willing to feel anger and willing to believe by themselves cannot produce the desired outcomes; in contrast, willing to perform voluntary actions such as raising my hand can produce the desired outcomes. If we wish to affect involuntary phenomena such as anger or belief, we will need to manipulate our stimuli; this gives us only at best very limited and crude control over the result, which is nowhere near sufficient for voluntariness.[3] Thus sentiments remain involuntary

[3] Feldman (2000, 2001) points out that there is one class of beliefs over which we have significant control, namely, beliefs about states of the world we can directly affect. For

 © Qu 2016

despite our having some weak indirect influence over them. All in all, for Hume beliefs are clearly involuntary, as evinced by his repeated statements to this effect.

1.2 The normative accountability of belief

It is equally uncontroversial that Hume holds our beliefs up to normative appraisal,[4] as his philosophy is replete with the endorsement and censure of various beliefs.[5] First, Hume repeatedly expresses his endorsement of some forms of causal inference, labelling them 'just' (THN 1.3.6.7, THN 1.3.13.3, THN 1.4.4.1); Hume also frequently labels beliefs, opinions and conclusions reasonable or unreasonable (THN 1.3.12.19; THN 1.4.3.1; THN 2.3.3.6, and THN 1.3.8.14; EHU 5.3). Furthermore, he states that inductive inferences provide *knowledge* (EHU 5.2), which implies their justification. Moreover, Hume is also unapologetic about criticising certain beliefs. For example, in discussing the sources of 'unphilosophical probability', Hume dismisses prejudices as 'errors' that we 'rashly form' (THN 1.3.13.7). In addition, Hume takes his argument on miracles to condemn 'arrogant bigotry' and 'superstitious delusions', freeing us from 'their impertinent solicitations' (EHU 10.2).

Nor is Hume shy about using deontological language in his epistemology, frequently making 'ought' claims in epistemological contexts. He

---

instance, I can cause myself to believe that the lights are on, simply by switching on the lights. He points out that this is not very epistemologically significant, because the large majority of beliefs that we do hold accountable do not fall under this category.

[4] A distinction is sometimes raised between the normative and the evaluative, the difference being that normative evaluations concern praise and blame (e.g. 'she is a good person') while evaluative ones do not (e.g. 'the cat has nice fur'). However, Hume does not seem to demarcate a realm of value that is independent of praise and blame, at least not where persons are involved. For example, many aesthetic qualities are thought to be evaluative rather than normative, but Hume treats them on a par with virtue and vice, which clearly concern praise and blame (THN 3.3.4.3); similarly with 'cowardice, meanness, levity, anxiety, impatience, folly', which Hume holds to be vices (EPM App 4.20), but are sometimes taken to be strictly evaluative in modern discussions. Thus, one cannot escape the inconsistency by claiming that Hume only holds beliefs subject to evaluative but not normative censure. Indeed, this strategy seems independently implausible anyway, since Hume frequently seems to praise and blame people on the basis of their beliefs, for example, he praises as wise those who proportion their belief to the evidence (EHU 10.4). Thanks to Robert Hopkins for helpful discussion on this issue.

[5] Of course, to endorse or censure a belief is simply to hold its believer accountable on the basis of this belief. Similarly, to praise a virtue or to condemn a vice is to praise or blame the agents who possess these qualities of the mind: Hume notes that the object of hatred or anger is 'a person or creature endow'd with thought and consciousness' (THN 2.3.2.6). For convenience I will speak of normatively appraising beliefs and other similar expressions, even though I mean normatively appraising an agent on the basis of her belief, etc.

 © Qu 2016

notes that hypotheses that pretend to discover the original principles of human nature 'ought at first to be rejected as presumptuous and chimerical' (THN Intro 8). He speaks of his general rules in THN 1.3.15 'by which we ought to regulate our judgment concerning causes and effects' (THN 1.3.13.1). In his 'Of Scepticism With Regard to Reason', Hume points out that 'we ought always to correct the first judgment, deriv'd from the nature of the object, by another judgment, deriv'd from the nature of the understanding' (THN 1.4.1.5), although this principle eventually proves problematic in engendering said scepticism regarding reason. Notably, in his Title Principle, Hume states: 'Where reason is lively, and mixes itself with some propensity, it ought to be assented to' (THN 1.4.7.11). In his discussion of miracles, Hume argues that 'where there is an opposition of arguments, we ought to give the preference to such as are founded on the greatest number of past observations' (EHU 10.16). Similarly, in discussing the possibility of a future state, Hume argues that 'If the cause be known only by the effect, we never ought to ascribe to it any qualities, beyond what are precisely requisite to produce the effect' (EHU 11.13). And in recommending his brand of mitigated scepticism, he points out that 'there is a degree of doubt, and caution, and modesty, which, in all kinds of scrutiny and decision, ought for ever to accompany a just reasoner' (EHU 12.24).

This is but a small selection of the many instances where Hume normatively appraises beliefs; it should, however, be sufficient to demonstrate in no uncertain terms Hume's holding that at least some beliefs are subject to normative appraisal.[6] Note that for the purposes of this paper, I assume the rejection of interpretations of Hume as a thoroughgoing sceptic; elsewhere, I argue for this claim more extensively, but to repeat the arguments here would take me too far afield.[7]

[6] It is fairly uncontroversial that Hume does hold beliefs up to normative appraisal. Nevertheless, it might be thought that his remark that 'false judgements … are commonly very innocent, and draw no manner of guilt upon the person who is so unfortunate as to fall into them' (THN 3.1.1.12) indicates that he in fact does not. However, a cursory examination of this passage will reveal that Hume is concerned here with *moral* appraisal, which of course says nothing about *epistemic* appraisal; Hume's point is merely that people are not held morally responsible for their false beliefs, which leaves open their being held epistemically responsible for them.

[7] I reject the reading of Hume as a thoroughgoing inductive sceptic in detail in my 'Hume's Positive Argument on Induction' (2014b); see also my 'Hume's Practically Epistemic Conclusions?' (2014a) and 'Hume's Internalism in EHU 12' (forthcoming) for a defence of interpretations of Hume that are not excessively sceptical in THN 1.4.7 and EHU 12 respectively. An interpretation of Hume as a thoroughgoing sceptic is in any case somewhat at odds with his self-professed mitigated scepticism in EHU 12.

 © Qu 2016

1.3 Ought implies Can

Before we go on to examine the plausibility of attributing 'Ought implies Can' to Hume, I will take a moment to deal with a *prima facie* objection to doing so. It might be thought that Hume cannot hold this principle given that he is perfectly willing to make judgments on moral responsibility despite his commitment to determinism. Just as we differentiated two senses of 'control' in discussing the involuntariness of beliefs, we can also distinguish two senses of 'can' with respect to 'Ought implies Can'. Hume's compatibilism does not entail that Hume cannot subscribe to 'Ought implies Can' *tout court*, but merely that he cannot subscribe to a certain *version* of 'Ought implies Can', where 'can' is given a libertarian reading as implying some sort of contra-causal agent-causation. (Call this 'Ought implies Libertarian-Can'.) Clearly Hume does not subscribe to this libertarian principle, given his denial of such forms of agent-causation (EHU 8.25). However, if we interpret 'can' as opposed not to necessity, but to constraint, we get a much more defensible reading of 'Ought implies Can'. We should read the relevant notion of 'Can' as concerning *voluntary* control: as mentioned previously, to say a person 'can X' in the sense of having voluntary control is just to say that he has '*a power of acting or not acting* [to do X], *according to the determinations of the will*' (EHU 8.23, emphasis in original). Thus understood, 'Ought implies Can' is broadly equivalent to saying that if a person ought to perform X, then X is suitably determined by her will. (Call this 'Ought implies Voluntary-Can'.) In what follows, when I discuss 'Ought implies Can', I mean 'Ought implies Voluntary-Can', unless otherwise stated. More specifically, the form of 'Ought implies Voluntary-Can' that is relevant to the puzzle of doxastic involuntarism is '*Epistemic*-Ought implies Voluntary-Can'; I take it that normative evaluations of beliefs fall under the epistemic (rather than, say, the moral or aesthetic) realm of normativity.

Having specified the relevant form of 'Ought implies Can', we can now evaluate the plausibility of attributing this principle to Hume. 'Ought implies Can', although traditionally viewed as a Kantian principle,[8] seems an intuitive constraint on normativity, as it seems unreasonable to fault someone for failing to perform an action of which she is incapable. Although commentators do not usually explicitly

[8] Reid also defends a version of 'Ought implies Can' in Chapter 6, Essay 6 of his *Essays on the Intellectual Powers of Man* (published in the years between Kant's First and Second *Critiques*), noting that a person should not be blamed for something that the person did not have the power to prevent.

 © Qu 2016

attribute 'Ought implies Can' to Hume, many implicitly hold him to it.[9] Other commentators do explicitly attribute 'Ought implies Can' to Hume. Ridge (2003, p. 172) does so on the basis of the following passage:

> ... false judgements ... are commonly very innocent, and draw no manner of guilt upon the person who is so unfortunate as to fall into them ... They extend not beyond a mistake of *fact*, which moralists have not generally suppos'd criminal, as being perfectly involuntary. (THN 3.1.1.12; SBN 459, emphasis in original)

Morris (2000, p. 105) also assumes Hume's endorsement of 'Ought implies Can', but only in passing. To support the claim that Hume endorses this principle, Wilson (1997, p. 202) points to Hume's dismissal of 'Pyrrhonian' scepticism due to the impossibility of anyone adhering to these standards of belief-formation: Hume remarks that 'nature is always too strong for principle' (EHU 12.23), and that 'Nature, by an absolute and uncontroulable necessity has determin'd us to judge as well as to breathe and feel' (THN 1.4.1.7).[10] Thus, there is at least a *prima facie* case for attributing 'Ought implies Can' to Hume.

[9] Price (1969, pp. 239–40) and Flew (1961, p. 98) believe Hume to be inconsistent in holding (a) and (b), which seems to presuppose 'Ought implies Can'. Owens (2000, p. 5) also holds that Hume's doxastic involuntarism rules out beliefs being subject to normative assessment. Similarly, Passmore (1980), Norton (1982, 1994, 2002) and Owen (1999), in recognising this puzzle (and going on to offer 'Prior Voluntary Action' as a solution to it), are taking (a) and (b) to be inconsistent positions, which again seems to presuppose 'Ought implies Can'. Although none of these commentators explicitly state the problem in terms of 'Ought implies Can', they clearly assume it in Hume in some form or another.

[10] Wilson attributes 'Ought implies Can' to Hume only incidentally, as his main aim is to establish that Hume subscribes to the stronger principle that 'Must implies Ought'. It is fairly easy to derive 'Ought implies Can' from 'Must implies Ought', using standard definitional principles and assuming that 'Ought implies Permissible':

1 Must X → Ought X

2 Ought X → Permissible X

3 Therefore, Must X → Permissible X

4 Taking the contrapositive of (3): ¬Permissible X → ¬Must X

5 By definition, ¬Permissible X ≡ Ought ¬X

6 Also by definition, ¬Must X ≡ Can ¬X

7 Substituting (5) and (6) into (4): Ought ¬X → Can ¬X,

which is 'Ought implies Can'.

 © Qu 2016

1.4 The puzzle

In the previous sections we saw that there is a case for attributing to Hume the following three claims: (a) All beliefs are involuntary; (b) Some beliefs are subject to normative appraisal; and(c) 'Ought implies Can'. As pointed out in §1, these three claims are jointly inconsistent. Here is a more detailed exposition of this inconsistency. Let us examine these claims with respect to my particular belief X, which is held normatively censurable:

(1) My believing X is involuntary (from a).

(2) If believing X is involuntary, then I cannot directly refrain from believing X.[11]

(3) I cannot directly refrain from believing X (from 1 and 2).

(4) 'Cannot implies not-Ought' (contrapositive of c).

(5) It is not the case that I ought to directly refrain from believing X (from 3 and 4).

(6) My belief X is subject to normative censure (b implies that at least one belief is subject to normative censure, and I assume it is mine for ease of exposition).

(7) If a belief is subject to normative censure, I ought to refrain from believing it.

(8) I ought to refrain from believing X (from 6 and 7).

As we can see, (5) and (8) straightforwardly contradict each other. Thus, it is inconsistent jointly to hold (a), (b), and (c). We thus have an interpretive puzzle (henceforth 'the puzzle'), given the initial plausibility of Hume's holding all three theses.[12] I will first examine the received 'Prior Voluntary Action' solution which can be read

[11] Of course, one might *indirectly* refrain from believing by undertaking some prior voluntary action (such as attending to certain evidence) that would lead to one's not forming the relevant belief. This will prove crucial to the 'Prior Voluntary Action' strategy. I examine this position shortly.

[12] This is, of course, a substantive philosophical puzzle as well as an interpretive one: all three claims (a), (b), and (c) are plausibly true but jointly inconsistent. This paper will focus primarily on the interpretive puzzle rather than the philosophical one—the question I primarily address is what consistent position (if any) Hume adopts, rather than which position is the most philosophically defensible. Of course, in scholarship the two cannot entirely be divorced: given considerations of charity, which I believe play a significant role in scholarship, philosophical defensibility does contribute to textual tenability.

 © Qu 2016

either as denying (a) or (b); I argue that, although initially plausible, both readings are in fact deeply problematic.

2. Prior Voluntary Action

2.1 Characterising the position

The ubiquitous solution to this puzzle in the Hume scholarship is to argue that beliefs, although in themselves involuntary, are the result of prior actions that are voluntary—call this position 'Prior Voluntary Action'.[13] Here I am construing 'actions' in a wide sense to include both physical and mental acts (e.g. the conjoining of ideas); I also take the non-performance of an act (for instance, failing to gather evidence) itself to count as an action. This seems to be a plausible way of resolving this puzzle, but it is worth investigating how it purports to do so. Given that the set {a,b,c} is inconsistent, which of the three premises is rejected by 'Prior Voluntary Action'? 'Prior Voluntary Action' can be read in two distinct ways. How exactly we understand it to solve the puzzle will depend on which reading we prefer. First, it might be seen as maintaining that beliefs are voluntary in a weakened derivative sense that is nevertheless relevant to 'Ought implies Can', that is, it can be seen as denying (a). Alternatively, it can be seen as arguing that beliefs are merely subject to derivative normative appraisal, which is not the type of normative appraisal that 'Ought implies Can' concerns; that is, 'Prior Voluntary Action' can be seen as denying (b).

According to the version of 'Prior Voluntary Action' that denies (a) (henceforth 'Voluntary Inheritance'), beliefs are derivatively voluntary insofar as they derive from voluntary actions, and 'Ought implies Can' extends to whatever is derivatively voluntary. Beliefs are both voluntary (in a weakened but nevertheless relevant sense) and epistemically assessable, and there is no contradiction of 'Ought implies Can'. According to the version of 'Prior Voluntary Action' that denies (b) (henceforth 'Normative Inheritance'), beliefs are not in themselves epistemically assessable because of their involuntariness, and 'Ought implies Can' is maintained. Nevertheless, they may possess derivative normative significance, insofar as they issue from actions for which we are responsible, because they are done voluntarily. 'Normative Inheritance' maintains that 'Ought implies Can' does not extend to

[13] In the modern philosophical debate on doxastic involuntarism, this position is defended by Alston (1988), Audi (2001, 2008), Huss (2009), and Leon (2002), among others.

 © Qu 2016

derivative normative significance; beliefs are both involuntary and do not possess intrinsic normative significance, in accordance with 'Ought implies Can', but we nevertheless speak loosely in attributing to them derivative normative significance in virtue of the actions that cause them.

One way of cashing out the difference between the two positions is through the notion of 'inheritance'. Let us roughly characterize the inheritance relation as follows: A 'inherits' X from B iff A's being X is determined (where determination is a non-symmetric relation) by B's being X. 'Voluntary Inheritance' takes beliefs to 'inherit' derivative *voluntariness* (in a sense relevant to 'Ought implies Can') from the prior voluntary actions that cause them; this allows for beliefs to be normatively assessable—recall that given 'Ought implies Can', normative assessability is impossible without voluntariness. Of course, this leaves open the basis on which these beliefs are subsequently normatively assessed; once they inherit voluntariness from their prior voluntary actions, their particular normative status might turn on factors entirely independent of these actions. Meanwhile, 'Normative Inheritance' takes beliefs to 'inherit' derivative *normative significance* (in a sense *not* relevant to 'Ought implies Can') from their prior voluntary actions; in contrast to 'Voluntary Inheritance', this does not leave open the basis on which beliefs are normatively assessable, since their normative status is directly inherited from and therefore determined by their prior voluntary actions.[14,15]

To sum up, instead of the inconsistent triad {a,b,c}, 'Voluntary Inheritance' accepts the following consistent triad:

(a') Some beliefs are derivatively voluntary,

(b) Some beliefs are subject to normative appraisal, and

(c') Ought implies (derivatively voluntary) Can.

[14] Note that both 'Voluntary Inheritance' and 'Normative Inheritance' read 'Ought implies Voluntary-Can' differently. I do not take either of them to reject (c), but rather to adopt different disambiguations of it.

[15] Note that it is implausible for a version of 'Normative Inheritance' to hold that beliefs inherit normative assessability in a loose way of speaking without also inheriting a particular normative status, because then their normative status would be underdetermined—not truly being assessable in their own right, beliefs will have to inherit their particular normative status from *something* voluntary if they are to have one at all.

© Qu 2016

'Normative Inheritance' instead accepts the following (different but equally consistent) triad:

(a) All beliefs are involuntary,

(b*) Some beliefs are subject to derivative normative appraisal, and

(c*) (Non-derivative) Ought implies Can.

2.2 Objecting to 'Prior Voluntary Action' in either form

One objection to attributing 'Prior Voluntary Action' in either form to Hume is the lack of textual support for Hume's taking prior voluntary actions to be essential to the evaluation of our beliefs.[16] Some textual evidence is put forth by Falkenstein (1997), Owen (1999), McCormick (2005), and Hickerson (2013). Falkenstein (1997, p. 33) and Owen (1999, p. 216) point to Hume's Endnote H to §9 in his first *Enquiry*. In this endnote, Hume seeks to explain differentials in reasoning ability between people and animals, and also between different people. Hume offers a variety of explanations; those important to Falkenstein's account are the ones involving prior voluntary actions, such as greater experience in reasoning by analogy (number 7 on Hume's list) and engaging in 'books and conversation' (number 9). However, this endnote fails to adequately substantiate 'Prior Voluntary Action'. For one, it is unclear that Hume is offering a normative rather than a merely descriptive explanation for differentials in reasoning ability. But even if we take him to be offering a normative explanation here, endnote H nevertheless fails to support 'Prior Voluntary Action'. Although a few of the considerations Hume offers are related to prior voluntary actions (as detailed above), the vast majority of them concern *involuntary* qualities (for example,

[16] Passmore (1980) was one of the earliest to offer this defence, and he does so at the end of his chapter without meaningful textual support. Passmore presents 'Prior Voluntary Action' as a position that Hume must hold in order to maintain 'an ethics of belief', rather than one Hume explicitly defends. Williams (2004, p. 274) defends 'Prior Voluntary Action' only in passing and without offering textual evidence. Norton (2002, p. 384) appeals to the fact that Hume exhorts us to proceed in our enquiries with a degree of doubt (EHU 12.4, EHU 12.24) as evidence that doubting is to some extent voluntary. But this is little evidence for 'Prior Voluntary Action': Hume also exhorts us to hold certain beliefs (e.g. beliefs that are proportionate to the evidence, in EHU 10.4), despite belief being involuntary. Indeed, it is difficult to see how doubt, as the diminishing of vivacity, could be voluntary if vivacity is an involuntary sentiment. As detailed in §1.1, there are of course voluntary actions that can engender doubt (such as considering the sorry state of our faculties), but this does not come close to establishing that Hume holds that beliefs are only assessable relative to their prior voluntary actions.

 © Qu 2016

superior memory, attention, and powers of observation; having a larger mind; being able to carry on a chain of consequences longer; greater accuracy and subtlety; a less hasty and narrow mind; and being less susceptible to bias). If we take Hume to be offering grounds for epistemic evaluation in endnote H, then it seems that he takes involuntariness to be irrelevant to epistemic evaluations. Thus, endnote H offers little evidence for 'Prior Voluntary Action'.

In a similar vein, McCormick (2005, pp. 7–9), citing Falkenstein (1997), appeals to Hume's account of general rules in THN 1.3.13, arguing that Hume is here offering a way for us to influence our beliefs by means of the application of general rules. Hickerson (2013) substantiates this line of argument by putting considerable weight on Hume's appeal to the 'second influence of general rules', by which 'we take a review of this act of the mind [of the first influence of general rules, involving overgeneralisation]', and 'find it to be of an irregular nature and destructive of all the most establish'd principles of reasoning; which is the cause of our rejecting it' (THN 1.3.13.12); Hickerson argues that this 'review' is an act of 'voluntary reflection'. However, this account is also unsatisfactory, because it fails to account for many epistemic judgments that Hume does indeed make. Indeed, in THN 1.3.13.3 – in the same section as the above passage – Hume argues that *immediate* causal inferences are 'just'; obviously, such immediate inferences do not involve any higher-order reflection involving general rules. Similarly, his claim that 'One who concludes somebody to be near him, when he hears an articulate voice in the dark, reasons justly and naturally; tho' that conclusion be deriv'd from nothing but custom' (THN 1.4.4.1) concerns an immediate inference which arises directly from custom, rather than from any higher-order reflection involving general rules. Note that it is not just that such inferences do not in fact involve any higher-order reflection, but that they *could not* involve any higher-order reflection—such beliefs are derived 'from nothing but custom', and so there is no higher-order reflection that could have been done (but was not) that could serve as a basis on which to evaluate these immediate beliefs. This is problematic for Hickerson's account.

Indeed, we can see that both THN 1.3.13 and endnote H ultimately fail to provide compelling textual evidence for 'Prior Voluntary Action', because they establish only that Hume thinks we can influence our beliefs through prior voluntary actions, which is, after all, an uncontroversial thesis; at a push, they also might be taken to show that Hume thinks that prior voluntary actions are *sometimes* relevant to

 © Qu 2016

normative evaluations, which is plausible but not particularly controversial. What 'Prior Voluntary Action' requires is evidence that Hume thinks the normative evaluation of *every* evaluable belief *inextricably rests* on the prior voluntary actions that produce them (since every belief is involuntary) and moreover is *only possible* with respect to these prior voluntary actions, and both Endnote H and THN 1.3.13 fall far short of providing evidence for this. Endnote H fails to establish this, because Hume there also cites involuntary characteristics as responsible for powers of reasoning, and so it cannot be evidence that he takes normative evaluations of beliefs to be only possible with respect to prior voluntary actions; pointing to Hume's appeal to general rules similarly falls short of the mark, because it fails to account for various forms of epistemic judgments that he makes which do not (and cannot) concern higher-order reflection involving general rules at all (e.g. THN 1.3.13.3; THN 1.4.4.1; THN 1.4.7.11). Therefore, these passages fail to adequately substantiate 'Prior Voluntary Action'.

Moreover, another objection to 'Prior Voluntary Action' in any form as a solution to the puzzle is that it fails to account for synchronic rationality. To clarify this objection, let us distinguish between diachronic and synchronic epistemic evaluations. Diachronic epistemic evaluations are evaluations of beliefs relative to an extended time period (e.g. evaluations regarding evidence-gathering; the development of intellectual virtues), while synchronic epistemic evaluations are evaluations of beliefs relative to a given point in time, without reference to anything that came before (e.g. evaluations of evidence-responsiveness at a given point in time).

One objection is that 'Prior Voluntary Action' is too weak to rescue an ethics of belief because 'Prior Voluntary Action' only explains diachronic evaluations of beliefs, since it claims that *temporally prior* actions are relevant to the epistemic evaluations of beliefs. However, Feldman (2000, 2001) argues that synchronic evaluations of beliefs are crucial to epistemic normativity:

> ... the control that we do have [over our beliefs] comes at the wrong point in the belief-forming process ... The real worry is that epistemic evaluations have to do with how we respond to evidence, and we don't have voluntary control over that. (2001, p. 83)

Chuard and Southwood (2009, pp. 610–1) echo this worry, and adduce more arguments in favour of this plausible intuition. They raise two reasons to think that synchronic norms are crucial to epistemic evaluation. The first reason is that it is an empirical fact that in

 © Qu 2016

order to satisfy diachronic norms, we must be able to satisfy synchronic norms. They cite evidence-gathering as an example: we can only effectively gather evidence if we satisfy the synchronic norm of being able to recognize evidence. The second reason they cite is that prior voluntary actions such as evidence-gathering are only epistemically valuable insofar as an agent satisfies synchronic norms such as being able to recognize evidence and being appropriately responsive to evidence: the most respectable evidence-gatherer would still be fundamentally irrational if she were unable to form appropriate beliefs in response to the gathered evidence.[17]

It is important to note that the converse does not hold: it is not true that in order to effectively satisfy synchronic norms, we require diachronic norms. One can be appropriately responsive to evidence even if one makes no extra efforts to gather evidence; similarly, one might naturally be appropriately responsive to evidence even if one makes no efforts to improve one's evidence-responsiveness over time. Moreover, it is not true that synchronic norms only have value insofar as the agent satisfies diachronic norms: being appropriately responsive to evidence still has value even if one does not satisfy diachronic norms such as evidence-gathering and developing intellectual virtues. In light of this, it seems that there is a sense in which diachronic norms are derivative on synchronic norms. Even if one does not accept this claim, the weaker claim that synchronic norms cannot merely be reduced to diachronic norms seems undeniable. This is of course not to say that an epistemology could satisfactorily leave out diachronic norms; plausibly, a satisfactory epistemology should accommodate both. An epistemology that leaves out synchronic norms is therefore incomplete; therefore 'Prior Voluntary Action' is unsatisfactory.

We can make the transition from a philosophical problem to an interpretive one by noting that some of Hume's epistemic exhortations concern synchronic rationality.[18] Notably, his famous remark that 'a wise man … proportions his belief to the evidence' (EHU 10.4)

[17] Note that there are both synchronic *and* diachronic norms governing evidence-responsiveness: synchronic norms governing evidence-responsiveness concern the agent's doxastic reaction to the available evidence when faced with it, while diachronic norms concern the development of intellectual virtues, which can influence synchronic evidence-responsiveness to some degree. My claim is that 'Prior Voluntary Action' can accommodate the latter, but cannot entirely account for the former.

[18] This is not to say that Hume is never concerned with diachronic rationality; certainly he would think that we have a responsibility to collect and attend to evidence. A good epistemology should account for both synchronic and diachronic rationality.

 © Qu 2016

seems to express epistemic approbation towards agents who synchronically respond appropriately to evidence. Of course, 'being wise' is not a punctuate act, since wisdom is a character trait, and is persistent through time. However, 'proportioning belief to the evidence' clearly seems a punctuate act; it is something that happens at a given point in time. In short, having a diachronic virtue (being wise) is dependent on synchronic acts of belief (proportioning belief to the evidence). Similarly, Hume's 'rules by which to judge of causes and effects' (THN 1.3.15) provide a normative framework for correctly responding to the available evidence concerning causal relations at a given time, rather than advocating the development of appropriate intellectual virtues or the gathering of evidence, and thus this framework also concerns synchronic rather than diachronic rationality. Moreover, what Garrett (1997) calls Hume's 'Title Principle' (THN 1.4.7.11) likewise regards synchronic rationality:

> Where reason is lively, and mixes itself with some propensity, it ought to be assented to. Where it does not, it never can have any title to operate upon us. (THN 1.4.7.11; SBN 270)

Here Hume is spelling out the conditions under which we should assent to a piece of reasoning at a particular time. The Title Principle is clearly an epistemic norm governing *synchronic* rationality, given that it concerns the conditions under which we should assent to reason at a given time. Since 'Prior Voluntary Action' cannot account for synchronic norms of rationality, 'Prior Voluntary Action' fails to accommodate the Title Principle. To sum up, insofar as Hume is often concerned with synchronic evaluations of beliefs, and insofar as 'Prior Voluntary Action' in either form only gives us at best diachronic evaluations of beliefs, 'Prior Voluntary Action' fails to do justice to Hume's epistemology, as it cannot accommodate an important class of his normative claims.[19,20]

[19] Another problem for 'Prior Voluntary Action' in either form is that of deviant causal chains. A belief could causally issue from a belief-policy, but in the wrong way, thereby blocking the 'inheritance' of either voluntariness or normative significance. Specifying what the 'right way' consists in is no trivial matter. I bracket such issues for the purposes of my paper.

[20] It may be objected that we only focus on the synchronic factors in epistemic evaluation because they are less remote and more relevant to our sentiments of approbation and blame, but diachronic considerations provide more reasoned support for our epistemic assessments. However, given that synchronic factors *are* relevant to our epistemic evaluations because they stir our sentiments of disapprobation and blame, it seems there is a clear sense in which synchronic factors are epistemically crucial, and so there is something that 'Prior Voluntary Action' fails to capture. As Hume says in EPM 4.21, 'But this, in the mean time, must be

 © Qu 2016

Besides these problems for any form of 'Prior Voluntary Action', 'Voluntary Inheritance' and 'Normative Inheritance' have their respective worries in being inconsistent with some of Hume's other theses. I will highlight these problems in the following two subsections. Even if this section has not yet convinced the reader that 'Prior Voluntary Action' is untenable as an interpretive solution to the puzzle, the following problems for its disambiguations should do so. I take it that inconsistency with a figure's other theses is a major strike against an interpretation, and so the following objections should prove decisive.

2.3 Objecting to 'Voluntary Inheritance'

In light of my distinction between the two forms of 'Prior Voluntary Action', it should be noted that proponents of this position generally seem to characterize 'Prior Voluntary Action' as 'Voluntary Inheritance'.[21] However, despite its popularity, I believe that 'Voluntary Inheritance' is unsatisfactory as an interpretation of Hume.

For one, the details of such an account are somewhat problematic. Being causally derived from a voluntary action confers derivative voluntariness only under certain conditions—for instance, my tripping over a rock was caused by my voluntarily walking on the road, but surely my tripping is not derivatively voluntary. The additional criteria include at the very least a stipulation that voluntariness is 'inherited' only when the subsequent involuntary result is a *foreseen* consequence of the earlier voluntary action; foresight seems a minimum condition for a consequence being derivatively under voluntary control. However, such a stipulation is difficult to implement with respect to 'Voluntary Inheritance' since in performing actions such as evidence-

---

allowed, that *sentiments* are every day experienced of blame and praise, which have objects beyond the dominion of the will or choice, and of which it behoves us, if not as moralists, as speculative philosophers at least, to give some satisfactory theory and explication'. A similar point can be made here. Just as a moral theory should account for our holding involuntary traits responsible since they stir our sentiments of blame and praise, an epistemic theory should account for our holding synchronic factors responsible since they stir our sentiments of blame and praise. Given that synchronic factors such as evidence-responsiveness significantly affect our epistemic praise and blame, they should count as epistemically relevant, just as involuntary characteristics count as morally relevant because they stir sentiments of moral praise and blame.

[21] C.f. Passmore (1980, pp. 175–6); Falkenstein (1997, p. 66); Owen (1999, p. 216), McCormick (2005, p. 7), and Williams (2004, p. 274). In correspondence, Hickerson indicates that he characterizes his position as 'Normative Inheritance' rather than as 'Voluntary Inheritance'.

 © Qu 2016

gathering, we do not typically foresee the resulting belief—and indeed, foreseeing the resulting belief when gathering evidence often results in the epistemically culpable phenomenon of confirmation bias![22]

We have seen in the previous section that there is little textual evidence that Hume thinks prior voluntary actions are relevant to the normative evaluation of beliefs. However, 'Voluntary Inheritance' lies in even more dire textual straits, as even the mere *concept* of the 'inheritance' of voluntariness is not one that Hume explores. 'Normative Inheritance' is much more plausible in comparison, as Hume recognizes many instances of normative 'inheritance': Hume takes actions and passions to 'inherit' a measure of reasonableness or unreasonableness from their associated beliefs (THN 3.1.1.12, THN 2.3.3.6); he also takes actions to 'inherit' their laudability and blameability from the underlying qualities that produce them (THN 3.3.1.4); moreover, he recognizes the 'inheritance' of epistemic justification via the 'founded on' relation in his treatment of induction (EHU 4.14).[23] Insofar as we take Hume to hold that beliefs 'inherit' one of either voluntariness or normative significance from the prior voluntary actions that produce them, it is textually far more likely that they 'inherit' normative significance.

Indeed, 'Voluntary Inheritance' actively seems textually problematic. Hume argues that natural abilities are as morally evaluable as moral virtues (THN 3.3.4.1), going on to note that they are not susceptible to influence from prior voluntary actions:

> Men have observ'd, that tho' natural abilities and moral qualities be in the main on the same footing, there is, however, this difference betwixt them, that the former are almost invariable by any art or industry; while the latter, or at least the actions, that proceed from them, may be chang'd by the motives of reward and punishment, praise and blame. (THN 3.3.4.4; SBN 609)

Here Hume notes that while moral qualities may be changed by suitable motivation, natural abilities are 'almost invariable by any art or industry'. 'Art or industry' might refer either to any prior actions on the part of the agent, or to other people's efforts to change the agent. Either reading rules out natural abilities being significantly influenced by the agent's prior actions. Obviously, the first reading

[22] Thanks to Don Garrett for raising this point in correspondence.

[23] At least he does so according to normative interpretations of Hume's argument on induction. See Millican (2002b) for a detailed examination of the structure of Hume's argument, with the transitivity of the 'founded on' relation made clear. Garrett (1997) and Owen (1999) disagree, reading Hume's argument descriptively. I defend a normative interpretation in my 'Hume's Positive Argument on Induction' (2014b).

 © Qu 2016

straightforwardly contradicts this option, but the latter reading rules it out as well. Since the agent's actions can be changed by reward and punishment, as Hume notes in THN 3.3.4.4 above, if the agent's actions could influence her natural abilities, then rewards and punishments could change natural abilities via affecting these actions. But clearly Hume thinks that rewards and punishments cannot change natural abilities. So it seems clear that Hume thinks that natural abilities cannot be meaningfully influenced by prior actions on the part of the agent, hence why these abilities are 'natural', rather than acquired.[24]

This renders it highly unlikely that natural abilities could be derivatively voluntary for Hume; since prior voluntary actions cannot meaningfully influence natural abilities, it seems implausible that any voluntariness could be 'inherited' from these actions—nor are there any other suitable candidates from which these natural abilities can 'inherit' voluntariness. This would mean that natural abilities are normatively assessable despite (a) being involuntary, and (b) failing to 'inherit' any voluntariness from prior voluntary actions. This certainly casts substantial doubt on 'Voluntary Inheritance' as an interpretation of Hume, since he seems to hold that an involuntary entity need not 'inherit' any derivative voluntariness in order to be normatively assessable.

2.4 Objecting to 'Normative Inheritance'

First, it should be noted that although Hume recognizes many instances of normative 'inheritance' (as previously noted in §2.3), there is no textual evidence that he thinks the normative significance of beliefs can be 'inherited' from the normative states of the actions that eventually produce them.

A more pressing objection is that 'Normative Inheritance' falls to a vicious regress in conjunction with Hume's normative framework. 'Normative Inheritance' holds that beliefs are derivatively subject to normative appraisal on the basis of the prior voluntary actions that cause them. However, Hume tells us that actions are not themselves the fundamental subjects of attributions of reasonableness or unreasonableness:

> Actions may be laudable or blameable; but they cannot be reasonable or unreasonable: Laudable or blameable, therefore, are not the same with reasonable or unreasonable. (THN 3.1.1.10; SBN 458)

[24] It may be objected that one can in fact improve one's natural abilities through prior actions, such as doing mental exercises to improve one's memory. Even so, Hume clearly seems not to think this is the case, which casts doubt on 'Voluntary Inheritance' as an interpretation, if not as a philosophical position.

 © Qu 2016

> ... false judgments may be thought to affect the passions and actions, which are connected to them, and may be said to render them unreasonable, in a figurative and improper way of speaking. (THN 3.1.1.12; SBN 459)

Hume clearly uses 'laudable' and 'blameable' as terms of moral approbation and blame (e.g. THN 3.1.1.3, THN 3.1.2.3, THN 3.3.1.11). And Hume typically uses 'reasonable' and 'unreasonable' as terms of epistemic approbation and blame: he describes philosophical probabilities as *reasonable* foundations of belief and opinion in THN 1.3.13.1; he describes the production of certain proofs in metaphysical subjects as 'reasonable' in THN 1.2.5.21; he claims that his account of necessity is 'reasonable' in THN 1.3.14.26; and he seeks to establish the principle of allegiance to government on more 'reasonable' principles in THN 3.2.9.2. Furthermore, he often uses 'reasonable' and 'unreasonable' hand-in-hand with other unambiguously normative terms, for example, he describes the fictions of the 'antient philosophy' as 'unreasonable and capricious' (THN 1.4.3.1). Hume therefore seems to be claiming that actions are subject to moral but not epistemic appraisal in THN 3.1.1.10 and THN 3.1.1.12.[25] And indeed, Hume seems to repeat this claim elsewhere in this section in claiming that 'reason can never immediately prevent or produce any action by *contradicting or approving* of it' (THN 3.1.1.10, emphasis added); given that epistemic normativity is surely reason-based, the impossibility of actions being contradicted or approved by reason seems to rule out their intrinsically bearing epistemic normativity.

But if this is so, then the work of 'Normative Inheritance' is not done; an explanation must be given in turn for how actions are derivatively epistemically appraised as reasonable or unreasonable if we are to accept that beliefs 'inherit' their reasonableness or unreasonableness from them.[26] Therefore, if one is to adopt 'Normative Inheritance', one

[25] It may be objected that Hume denies a distinction between epistemic and moral normativity in THN 3.3.4, wherein he argues against there being a substantive distinction between natural abilities and moral virtues. However, Hume is not denying that there is a substantive distinction between moral and epistemic normativity; rather, he is merely making the point that intellectual qualities can be morally assessed on the basis of their usefulness and agreeableness to the possessor and to others, and so it makes little sense to call benevolence a moral virtue, but not a profound genius. This leaves it open that a profound genius can also be *epistemically* assessed on the basis of (say) its truth-conduciveness, and also count as an *epistemic* virtue. Thus, these passages are perfectly consistent with Hume thinking there to be a substantive distinction between moral and epistemic normativity.

[26] Could it be that beliefs 'cross-inherit' reasonableness or unreasonableness from the laudability or blameability of the actions that cause them, therefore halting the potential regress? I find the suggestion that something could 'inherit' a different kind of normative appraisal in

 © Qu 2016

must search deeper for the source of the unreasonableness of the prior voluntary action itself. For Hume, all voluntary actions are the products of beliefs and desires. So we have two *prima facie* candidates for this source: the doxastic component (i.e. the beliefs) or the affective component (i.e. the passions) that jointly produce the prior voluntary action.

If it is claimed that the fundamental subject of epistemic normativity is the doxastic component, then it should become immediately obvious that we set ourselves upon a vicious regress. For we are claiming that the initial involuntary belief is derivatively unreasonable on the basis of the prior voluntary action that causes it, and this prior voluntary action is itself derivatively unreasonable on the basis of the beliefs that produce it; yet these beliefs are themselves involuntary! We have come no closer to solving the puzzle of how Hume attributes normative accountability to involuntary beliefs. If we pursue this line of defence any further, we embark upon a vicious regress—beliefs are derivatively subject to normative appraisal in virtue of other beliefs, which are derivatively subject to normative appraisal because of other beliefs, and so on.

However, to turn to the affective component (that is, the end-setting passion or passions) would do no better. Recall that THN 3.1.1.12 makes clear that the passions, like actions, 'inherit' their reasonableness or unreasonableness from their associated beliefs; similarly in THN 2.3.3.6:

> Where a passion is neither founded on false suppositions, nor chooses means insufficient for the end, the understanding can never justify nor condemn it … a passion must be accompany'd with some false judgment,

---

this way a doubtful one. First, the notion of 'cross-inheritance' seems problematic, as different types of normativity do not seem to interact. For instance, one cannot 'inherit' prudential blame from moral blame: say John performs a cruel act A, which causes a state of affairs such that he is able to perform an action B that benefits his long-term interests—his performing action B is prudentially blameless despite proceeding from a morally blameworthy action, and there is no cross-inheritance of normativity. More importantly, note that the moral status of A has *no bearing* on the prudential status of B: his performing B is prudentially blameless, and his failing to perform B would have been prudentially blameworthy, regardless of whether A was moral or immoral. Moreover, from a textual point of view, the only cases of normative 'inheritance' Hume recognizes involve the direct 'inheritance' of the same sort of normative appraisal: see THN 3.1.1.12, given above, which involves the 'inheritance' of reasonableness or unreasonableness; or the 'inheritance' of laudability and blameability by actions from underlying qualities (THN 3.3.1.4); or the 'inheritance' of epistemic justification via the 'founded on' relation in his discussion of induction (EHU 4.14). Indeed, Hume explicitly denies the cross-inheritance of moral normativity from epistemic normativity in THN 3.1.1.12: here he rules out the possibility of moral deficiency stemming from errors in belief, which 'are so far from being the source of all immorality'; this renders it unlikely that the converse could hold, viz. that epistemic merit or demerit could derive from moral merit or demerit.

 © Qu 2016

> in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment. (THN 2.3.3.6; SBN 416)

Here, Hume clearly states that passions are only derivatively (also only figuratively and improperly said to be) unreasonable, that is, unreasonable in virtue of the unreasonableness of their associated beliefs. But as we have seen, delegating normativity to a belief brings us no closer to solving the puzzle of how an involuntary belief can be held up to normative censure in the first place.

To sum up this argument, 'Normative Inheritance' argues that the normative significance of a belief is derived from the normative appraisal of the prior voluntary actions that produce it—beliefs 'inherit' their epistemic significance from these prior voluntary actions. However, these prior voluntary actions are not themselves the fundamental subjects of attributions of reasonableness and unreasonableness, and are only derivatively so (THN 3.1.1.12). If it is claimed that attributions of reasonableness and unreasonableness toward these prior voluntary actions are 'inherited' from their doxastic component, we get a straightforward regress; however, to turn to the affective component is no better, as passions are themselves only derivatively rendered reasonable or unreasonable via their associated beliefs (THN 2.3.3.6; THN 3.1.1.12), leading to a vicious regress all the same.[27]

Could actions 'inherit' their reasonableness or unreasonableness from epistemic character traits, in the same way that they arguably 'inherit' laudability and blameability from moral character traits (THN 3.3.1.4)? This possibility is unlikely, for two reasons. First, Hume claims that actions 'inherit' their unreasonableness from associated *beliefs* (THN 3.1.1.12). Secondly, epistemic character traits are themselves involuntary; they are not determined by the will, and so they are beyond our voluntary control. Proponents of 'Normative Inheritance' typically maintain 'Ought implies Can', which is why they argue that involuntary beliefs only derivatively possess their normative statuses. To argue that the ultimate bearer of normative

[27] Note that the argument goes through even when the relevant 'prior voluntary action' is the non-performance of a particular action, since the non-performance of an action is an action in the relevant Humean sense. Take the example of my forming a belief regarding the weather tomorrow. Instead of checking weather reports, I instead stare blankly at a wall. The belief is culpable because the entire set of my prior actions (in this case staring blankly at the wall) is culpable in virtue of failing to contain the action 'check the weather report' (or relevantly similar actions). This set of prior actions, because voluntary, are produced by beliefs and desires, and so the argument I raised goes through the same.

 © Qu 2016

significance is itself involuntary would be to begin by denying 'Ought implies Can', thus resolving the puzzle of how involuntary beliefs are held normatively accountable in the first place; the solution to this puzzle would then be the denial of 'Ought implies Can', exactly as I go on to argue. That is, one could consistently hold both 'Normative Inheritance' and also deny 'Ought implies Can', but then it would be the latter rather than the former that resolves the puzzle. Therefore, 'Normative Inheritance' cannot maintain that the ultimate sources of epistemic significance are involuntary epistemic character traits, if it is to be taken as a resolution of the puzzle.

One might object to my regress argument by arguing that Hume does not use 'reasonable' and 'unreasonable' as terms of epistemic normativity in THN 3.1.1. I lack the space here to fully defend this interpretive claim, although I say more on this elsewhere;[28] in any case, the same conclusion can be reached via a different route, independently of how one reads Hume's usage of 'reasonable' and 'unreasonable'.

Hume emphasizes that actions cannot be the fundamental basis on which a person is judged morally vicious or virtuous because they lack sufficient durability, and so cannot influence our sentiments:

> If any *action* be either virtuous or vicious, 'tis only as a sign of some quality or character. It must depend upon durable principles of the mind, which extend over the whole conduct, and enter into the personal character. Actions themselves, not proceeding from any constant principle, have no influence on love or hatred, pride or humility; and consequently are never consider'd in morality. (THN 3.3.1.4; SBN 575, emphasis in original)

Insofar as we take epistemic normativity to require an affective response as well,[29] actions, because of their transience, cannot be the fundamental basis on which a person is judged epistemically reasonable or unreasonable. By parallel to the moral case, they must therefore serve as an indication of a durable quality if they are to play a role in epistemic evaluation. Recall also that for 'Normative Inheritance', the basis on which a person is epistemically judged must be voluntary if it is to be a viable solution to the puzzle. So we have two criteria that the basis on which we are epistemically judged must satisfy if

[28] See my 'Hume's Practically Epistemic Conclusions?' (2014a).

[29] The details of the epistemically relevant sentiments are somewhat orthogonal to my purposes in this paper, but 'curiosity, or the love of truth' (THN 2.3.10) seems a likely candidate, as does an uneasiness at discovering one's beliefs are false (thus motivating one to ensure the truth of one's beliefs).

 © Qu 2016

'Normative Inheritance' is to be viable: this basis must be (a) durable and (b) voluntary.

The problem is that anything that is sufficiently durable to play this normative role will have to be involuntary. Actions and acts of the imagination are free but transient, for example, while character traits, the underlying operations of the imagination, and the functioning of our faculties are durable but involuntary, since none is determined by the will. Whatever we take to be the basis on which we are epistemically judged – whether this basis is a character trait, or a dispositional belief,[30] or a belief-forming operation such as custom – will only be sufficiently durable to play this normative role if it is involuntary. And this means that the fundamental question of how we can attribute epistemic normativity to an involuntary quality remains unanswered. In the end, it seems that the only solution to the fundamental problem is to reject 'Ought implies Can'.[31]

Given the dearth of compelling textual evidence for 'Normative Inheritance', as seen in §2.2, there seems no strong reason to attribute to Hume a position that proves inconsistent with his other theses. I therefore reject 'Normative Inheritance' as a viable interpretation.

3. Denying 'Ought implies Can'

In this section, I will give two arguments for the claim that Hume does not subscribe to 'Ought implies Can', and therefore can consistently hold both (a) and (b).[32,33] First, this claim is the least textually

[30] Loeb (2002) argues that Hume is fundamentally committed to a dispositional account of beliefs, although his official theory treats them as occurrent. For a dissenting view, see Marusic (2010). In any case, the issue is largely orthogonal to my purposes in this paper.

[31] Of course, the arguments in this section do not preclude beliefs from 'inheriting' normative significance from actions in some (but crucially not all) cases. Indeed, I find it plausible that in certain diachronic epistemic evaluations, beliefs are partly evaluated in light of the prior voluntary actions that produce them. But this cannot explain how involuntary beliefs can be held accountable, since actions cannot be the fundamental subjects of epistemic normativity, and must inherit epistemic significance from other beliefs. Whatever the details of Hume's epistemology, this puzzle is ultimately resolved by denying 'Ought implies Can'.

[32] My position agrees with that of Vitz (2009), who argues that for Hume, beliefs are not subject to voluntary control but are nevertheless subject to normative appraisal, also citing Hume's treatment of character traits.

[33] For a defence of denying 'Ought implies Can' with respect to epistemic obligations in the modern debate on doxastic involuntarism, see Wolterstorff (1997), who compares epistemic oughts to 'paradigm' oughts, and Feldman (1988, 2000, 2001), who gives two differing explanations. In his 1988 paper, he argues that we can have obligations that we are not capable of carrying out, for example our obligation to pay our mortgage despite not

 © Qu 2016

substantiated of the three jointly inconsistent claims (a), (b), and (c); secondly, Hume clearly holds involuntary characteristics to be subject to praise and blame in the moral case, therefore removing any presumption that he should not do the same in the epistemic case.

3.1 Weakness of textual support

Of the three jointly inconsistent statements (a), (b), and (c) attributed to Hume, 'Ought implies Can' seems to have the weakest textual support by far; if we are looking to dismiss one of the three, this seems the best candidate. That he holds beliefs to be involuntary is substantiated by explicit textual evidence in which he forthrightly and unambiguously makes this claim, as we saw in §1.1. Meanwhile, that he holds beliefs up to normative appraisal is substantiated by numerous instances throughout his philosophical work, as we saw in §1.2. In contrast, the claim that Hume subscribes to 'Ought implies Can' is clearly the weak link of the three.

Wilson's suggestion that Hume's dismissal of extreme forms of scepticism derives from this principle is substantially weaker than the evidence we have seen for the other two claims, both in terms of quality and quantity. The main textual support that Wilson adduces is Hume's citing the involuntariness of inductive beliefs in his discussions of scepticism (THN 1.4.1.7; EHU 12.23), but this certainly falls far short of Hume endorsing 'Ought implies Can' in any form; indeed, Hume might be read as giving a *practical* (rather than normative) dismissal of sceptical worries by pointing out their irrelevance to our everyday practices, as Beebee (2006, p. 69) argues.[34]

---

having any money; the claim is that epistemic obligations are of this sort. Feldman (2000, 2001) later rejects this view in favour of one close to Wolterstorff's; he claims that epistemic obligations are more like 'role' oughts, which 'describe the right way to play a certain role' (p. 676); examples are the obligations associated with being a teacher, or a commander, or indeed a believer. Denials of 'Ought implies Can' which also extend to the moral case are given by Sinnott-Armstrong (1984), Stocker (1971), and Saka (2000), although not in the context of the debate regarding doxastic involuntarism. Ryan (2003) does the same, but while explicitly discussing doxastic involuntarism; indeed, she argues that Feldman (2000, 2001) should also deny the principle with respect to morality, given that we all play roles as 'moral agents', just as we play roles as 'believers'. Chuard and Southwood (2009) take themselves to maintain 'Ought implies Can', but essentially argue against there being a relevant notion of 'voluntary-can'.

[34] Although I read EHU 12.23 as giving a merely practical dismissal of scepticism, I argue in my 'Hume's Internalism in EHU 12' (forthcoming) that Hume also gives a robust epistemic refutation of excessive scepticism elsewhere.

 © Qu 2016

Ridge's (2003, p. 172) textual support seems more promising:

> ... false judgements ... are commonly very innocent, and draw no manner of guilt upon the person who is so unfortunate as to fall into them ... They extend not beyond a mistake of *fact*, which moralists have not generally suppos'd criminal, as being perfectly involuntary. (THN 3.1.1.12; SBN 459, emphasis in original)

However, while Hume clearly agrees with the 'moralists' that false beliefs are not morally reprehensible, it is doubtful that he agrees with them on the *basis* of this conclusion, viz. that they are involuntary. This explains why in THN 3.1.1.12 above, Hume unreservedly endorses the claim that false beliefs are not morally blameable while being careful to put the 'involuntariness' basis for this claim in the mouths of 'moralists'—note also the hedging terms 'commonly' and 'generally suppos'd'. Far from endorsing the thought that beliefs could not be morally blameable because they are involuntary, Hume wields this argument as an *ad hominem* attack against his rationalist opponents, who typically accept 'Ought implies Can' while maintaining that moral turpitude derives from false beliefs. Indeed, were Hume to agree with the moralists on this matter, this would create a tension with his claim in THN 3.3.4.3 that natural abilities are subject to moral appraisal despite being involuntary.[35]

In the absence of textual substantiation for Hume's subscription to 'Ought implies Can', the major motivation for attributing this principle to Hume seems to be its commonsensical status, but Hume's philosophy is frequently revolutionary. Furthermore, the fact that Hume never seems to recognize a tension between his subscriptions to theses (a) and (b) seems to provide at least a *prima facie* interpretive reason not to attribute 'Ought implies Can' to him. On the whole, if we are to interpret Hume as rejecting one of the three claims, 'Ought implies Can' seems the best bet.

3.2 A parallel to moral normativity

The second and most powerful objection to the claim that Hume subscribes to 'Ought implies Can' points to a peculiarity in that Hume's moral evaluation of qualities of the mind is not typically considered problematic, despite their being just as involuntary as beliefs. If we take this attribution of normativity by Hume to be textually unproblematic, then we are implicitly taking Hume to reject 'Ought implies Can' with respect to qualities of the mind in the moral case,

[35] This claim will be discussed more extensively shortly.

 © Qu 2016

which lends much plausibility to the notion that he rejects 'Ought implies Can' in the epistemic case as well.[36]

Hume clearly holds that we are legitimately held morally accountable on the basis of the involuntary qualities of the mind:

> In general, we may observe, that the distinction of voluntary or involuntary was little regarded by the ancients in their moral reasoning; where they frequently treated the question as very doubtful, *whether virtue could be taught or not*? They justly considered, that cowardice, meanness, levity, anxiety, impatience, folly, and many other qualities of the mind, might appear ridiculous and deformed, contemptible and odious, though independent of the will. Nor could it be supposed, at all times, in every man's power to attain every kind of mental, more than of exterior beauty. (EPM App 4.20; SBN 321-2, emphasis in original)

Here Hume approvingly remarks that the ancient philosophers *justly* condemned involuntary characteristics. Whether something is normatively censurable is not, therefore, a function of its attainability—in other words, ought does not imply can! Hume goes on to say:

> Philosophers, or rather divines under that disguise, treating all morals, as on a like footing with civil laws, guarded by the sanctions of reward and punishment, were necessarily led to render this circumstance, of *voluntary* or *involuntary*, the foundation of their whole theory. Every one may employ *terms* in what sense he pleases: But this, in the mean time, must be allowed, that *sentiments* are every day experienced of blame and praise, which have objects beyond the dominion of the will or choice, and of which it behoves us, if not as moralists, as speculative philosophers at least, to give some satisfactory theory and explication. (EPM App 4.21; SBN 322, emphasis in original)

This is echoed in the *Treatise*:

> Those, who represent the distinction betwixt natural abilities and moral virtues as very material, may say, that the former are entirely involuntary, and have therefore no merit attending them, as having no dependence on liberty and free will. But to this I answer, *first*, that many of those qualities, which all moralists, especially the antients, comprehend under the title of moral virtues, are equally involuntary and necessary, with the qualities of the judgment and the imagination … *Secondly*, I wou'd have any one give me a reason, why virtue and vice may not be involuntary, as well as beauty and deformity. These moral distinctions arise from the natural distinctions

[36] Admittedly, Hume clearly draws a distinction between epistemic and moral normativity, as noted in §2.4, which might be thought to compromise the strength of this analogy. But both types of normativity appear to share a common logic (e.g. the equivalence between 'X is not obligatory' and 'not-X is permissible' in the relevant senses), and this makes it odd to accept 'Ought implies Can' with respect to epistemic rationality but not with respect to morality.

 © Qu 2016

> of pain and pleasure; and when we receive those feelings from the general consideration of any quality or character, we denominate it vicious or virtuous. (THN 3.3.4.3; SBN 609-10, emphasis in original)

Again, here Hume is dismissing the view that voluntariness is required for moral responsibility, and he is happy to hold involuntary characteristics morally responsible—indeed, he holds *only* involuntary characteristics culpable in the moral evaluation of agents, since the basis on which we are morally assessed (i.e. qualities of the mind) are beyond our voluntary control. Hume argues that what determines matters of moral responsibility are the moral distinctions which arise from our natural sentiments of pain and pleasure,[37] and these natural sentiments are not directly sensitive to considerations of voluntariness.[38] The voluntariness of qualities of the mind is therefore irrelevant to their normative status, according to Hume. Why then should we think that he sees things differently in the epistemic case? Perhaps epistemic worth is seen by Hume as akin to aesthetic and moral worth, and is best considered a kind of *beauty*;[39] such evaluations of beauty do not rest on considerations of voluntariness, and involuntary epistemic beauty is as pleasing to our sentiments as voluntary epistemic beauty.

However, there seems to be a bit of a puzzle here, for denying the intuitive 'Ought implies Can' is a difficult bullet to bite. If we read Hume as claiming that voluntariness is utterly irrelevant to normative evaluations, it seems we commit him to a counterintuitive position, with ludicrous consequences such as attributing equal moral blame to the man who accidentally kills someone as to the man who wilfully and knowingly commits murder. Of course, this is not Hume's position; indeed, he rejects the doctrine of 'liberty or chance' (that is, the denial of determinism) partly on the grounds that it fails to account for the fact that we tend to be held responsible for voluntary but not accidental actions, since Hume takes the denial of determinism to undermine the connection between an action and the underlying durable character of the person who performed it (THN 2.3.2.6).

[37] These natural sentiments are corrected and systematized first, which distinguishes Hume's position from the crudest forms of moral subjectivism.

[38] Of course, considerations of voluntariness affect our moral sentiments in more indirect ways, particularly with respect to actions. I discuss this shortly.

[39] Hume notably speaks of moral beauty in THN 3.1.1.21, THN 3.2.1.8, THN 3.2.2.1, THN 3.2.6.4, EPM 1.9, EPM App 1.13, and EPM App 4.20. He also makes reference to 'beauty and deformity in action, composition, and external objects' (THN 2.1.1.3); 'beauty and deformity in action' most likely refers to virtue and vice respectively, as Loeb (1977) points out.

 © Qu 2016

We can make the bullet of denying 'Ought implies Can' easier to chew by pointing out that voluntariness does play some role in the normative evaluation of actions but not beliefs or qualities of the mind, therefore going some way to explaining the intuitiveness of 'Ought implies Can'. But what reason could there be for drawing this arbitrary-seeming distinction? I propose that the answer centres on the evidential role that voluntariness plays with respect to actions. Recall that actions are not the fundamental basis on which a person is epistemically assessed, and play only an informative role in revealing underlying character: [40]

> If any *action* be either virtuous or vicious, 'tis only as a sign of some quality or character ... Actions are, indeed, better indications of a character than words, or even wishes and sentiments; but 'tis only so far as they are such indications, that they are attended with love or hatred, praise or blame. (THN 3.3.1.4-5; SBN 575, emphasis in original)

Voluntariness is relevant to our (derivative) appraisal of actions because, *ceteris paribus*, voluntary actions usually accurately reflect underlying character, while involuntary actions usually do not accurately reflect character. For example, we should not morally blame someone who involuntarily spasms and steps on our foot, because we recognize that this action did not proceed from his character; in contrast, it would be permissible to blame someone who deliberately trod on our foot, as this action would accurately paint an unflattering picture of his underlying character.[41]

[40] To accurately reflect underlying character, an action will have to proceed from it: as we cannot directly observe character, we are only licensed to make inferences about it via its visible products, viz. the actions it produces; these actions are only causally linked to character traits by being produced by them, and so the only way for an action to accurately reflect a character trait is by proceeding from it. However, it is not sufficient for an action to be used to evaluate a character trait that it proceeds from, but does not accurately reflect, that trait. Take the example of a man who performs an action out of greed but successfully hides the motive for this action. In this case, we are not in an epistemic position to subjectively judge him greedy on the basis of his apparently non-greedy action. Of course objectively speaking, we would judge him on the basis of his greedy character, but not on the basis of his action, since actions are too transient to be properly used as a basis for evaluation. Insofar as actions are subjectively relevant to normative evaluations, we are not licensed to judge the man to be greedy on the basis of his apparently non-greedy action. In short, an action must proceed from *and* accurately reflect underlying character to adequately signify underlying character.

[41] It might be counterintuitive to some to think that a voluntary action that did not accurately reflect a durable character trait of the agent—perhaps because it merely issued from a transient quality—would not count towards the moral evaluation of this person. Nevertheless, this seems to be Hume's view, as he states it in THN 3.3.1.4: if an action is not 'a sign of some quality or character', then it cannot be 'either virtuous or vicious'. So a

 © Qu 2016

This point can be verified by observing that where involuntary actions accurately reflect underlying character, we are more inclined to hold these actions accountable: if I am of a mean character and am so used to insulting others that I produce abusive comments automatically and involuntarily, I would be held equally (or more) accountable than if these comments had been voluntary. Therefore, it seems that the involuntariness of an action serves to excuse an agent only if it entails that this action is unreflective of underlying character.[42] Where involuntariness entails that an action is unreflective of underlying character, we are inclined not to hold the agent responsible for this action; conversely, where an involuntary action is reflective of underlying character, we are inclined to blame this action, as Hume notes:

> By the intention we judge of the actions, and according as that is good or bad, they become causes of love or hatred. But here we must make a distinction. *If that quality in another, which pleases or displeases, be constant and inherent in his person and character, it will cause love or hatred independent of the intention: But otherwise a knowledge and design is requisite, in order to give rise to these passions.* (THN 2.2.3.3-4; SBN 348-9, emphasis added)

---

voluntary action that did not accurately reflect a durable character trait would not properly count towards the moral evaluation of the agent. This is Hume's explanation for why 'repentance wipes off every crime' (THN 2.3.2.7); sincere repentance demonstrates that the past action no longer accurately reflects the agent's (now reformed) character.

[42] This is essentially the 'epistemic' solution to the problem of moral luck. (See Bernard Williams 1981 for a classic statement of this puzzle.) Briefly, the problem of moral luck may be stated as follows: how can we be held morally accountable on the basis of factors beyond our control? Take an example: driver 1 fails to check his brakes regularly and consequently runs over a child, while driver 2 is equally negligent but luckily never encounters any pedestrian children. We judge driver 1 more harshly than driver 2; yet the only difference between them (that is, a child's running in front of driver 1's car) was beyond their control. The epistemic solution to this puzzle (propounded by Richards 1986, Thomson 1993, and Rescher 1993, among others) argues that drivers 1 and 2 are equally morally accountable since they have the same characters, but we are not in an epistemic position to know this, and therefore presume the unlucky driver to have a worse character. In short, it is characters that we normatively assess, with actions playing a merely evidential role in allowing us to discern underlying character. Latus (2000) argues that this solution fails because it leaves unanswered the further question of how we can be morally judged on the basis of our involuntary characters; but as we have seen, Hume would at this point head off the potential regress by arguing that voluntariness is irrelevant to moral evaluations beyond playing a merely evidential role with respect to actions; that our character is beyond our control is not morally relevant, given that these characters are equally apt to excite love and hatred, pride and humility whether voluntary or not.

 © Qu 2016

Indeed, in the *Enquiry* Hume argues that voluntary freedom (viz. 'liberty') is crucial to morality precisely because it allows for actions to have evidentiary value regarding our durable character traits:

> It will be equally easy to prove, and from the same arguments, that liberty, according to that definition above mentioned, in which all men agree, is also essential to morality, and that no human actions, where it is wanting, are susceptible of any moral qualities, or can be the objects either of approbation or dislike. For as actions are objects of our moral sentiment, so far only as they are indications of the internal character, passions, and affections; it is impossible that they can give rise either to praise or blame, where they proceed not from these principles, but are derived altogether from external violence. (EHU 8.31; SBN 99)

On the other hand, voluntariness plays no role in moral appraisal beyond this evidential one, and therefore Hume can reject 'Ought implies Can' with respect to qualities of the mind, since they are the fundamental basis on which we are morally assessed and do not serve as mere signs of some underlying quality.

In the epistemic case, beliefs are either the fundamental basis on which we are epistemically assessed, or they—like actions in the moral case—merely play an informative role insofar as they reflect some underlying disposition of the agent. If the former is the case, then voluntariness plays no role in normative evaluation, given that beliefs would not serve as mere signs of some underlying quality.

On the other hand, if beliefs play only an informative role in epistemic evaluation, voluntariness would equally play no role in the normative evaluation of beliefs, given that an involuntary belief typically will nevertheless accurately reflect underlying epistemic character. In this respect, involuntary beliefs are like involuntary actions, such as my automatically insulting others due to my unpleasant character; these actions, like beliefs, proceed directly from underlying (epistemic or moral) character without going through a volitional process, and thus their involuntariness is not apt to mitigate blame or approbation, since they are excellent signs of this underlying character. Contrast this with innocent involuntary actions such as accidentally tripping and treading on someone's foot; such actions do not proceed from underlying character at all, and give one no reason to think poorly of this character. Of course, involuntary beliefs do not always accurately reflect underlying character, for example, when resulting from an uncharacteristic error in judgment. The point is that involuntary beliefs *usually* accurately reflect underlying epistemic

 © Qu 2016

character; on the other hand, involuntary actions are *usually* unreflective of underlying moral character. The involuntariness of belief does not in itself compromise the evidential value of the belief with respect to underlying epistemic character; if evidentiary value is compromised, it will typically be for reasons unrelated to voluntariness (e.g. the agent being overtaken by emotion, causing uncharacteristic errors in judgment). Thus, the involuntariness of beliefs does not compromise their evidential value with respect to underlying epistemic character. Therefore, considerations of evidential role explain why Hume would consider voluntariness so irrelevant to the normative evaluation of qualities of the mind and beliefs but not to the normative evaluation of actions.

All in all, it should be fairly clear from Hume's own words that he rejects 'Ought implies Can' with respect to qualities of the mind, and we have a principled reason for extending this rejection to beliefs, while maintaining that voluntariness is relevant to the evidential value of actions in normatively assessing the underlying characters that produce them. Therefore, Hume does not commit himself to an inconsistency in holding that beliefs are involuntary while at the same time subjecting them to normative censure. This interpretation is, unlike both forms of 'Prior Voluntary Action' examined earlier, not only consistent with Hume's theses, but also consonant with them in having a neat parallel to his moral account; on these grounds, I take it to be a much more plausible interpretation.

4. Parallels with moral responsibility and determinism

In §1.3, I pointed out that the above discussion is to a large degree independent of Hume's position on moral responsibility and determinism, as the two discussions involve different readings of 'Ought implies Can'. Nevertheless, it may be illuminating to draw out briefly some parallels between the two. To facilitate this comparison, let us disambiguate the two relevant readings of 'Ought implies Can':

Ought implies Libertarian-Can: If a person ought to perform X, this implies that she can perform X in the libertarian sense of having the capacity of agent-causation to perform X; she is able to operate in a non-deterministic, self-caused manner in performing X.

Ought implies Voluntary-Can: If a person ought to perform X, this implies that X is under her voluntary control: X is determined by her will.

 © Qu 2016

In Hume's discussion of determinism and moral responsibility, he faces three jointly inconsistent positions:

(x) Our actions are determined and therefore beyond libertarian control;

(y) Our actions are subject to normative censure;

(z) 'Ought implies Libertarian-Can'.

Compare this with the three jointly inconsistent positions Hume faces in the case of doxastic involuntarism:

(a) Our beliefs are beyond our voluntary control;

(b) Our beliefs are subject to normative censure;

(c) 'Ought implies Voluntary-Can'.

The parallels between the two should be immediately obvious; both {abd} and {xyz} take an element that is subject to normative censure and beyond our control in a certain way, along with a principle that dictates that normative censure is subject to that form of control.[43]

[43] The parallel between {xyz} and {abc} is not entirely direct since they concern different notions of control (voluntary versus libertarian). I think the comparison between these two sets is nevertheless the most edifying for my purposes, and {xyz} is a case Hume explicitly discusses, which is why I have focused on {xyz} in the main text. We can briefly examine the directly parallel sets (that are still inconsistent) in this footnote. Examine the set {a'bz}:

(a') Our beliefs are beyond libertarian control;

(b) Our beliefs are subject to normative censure;

(z) 'Ought implies Libertarian-Can'.

Hume does not directly discuss this case, since his discussion on free will and determinism centers around actions (rather than beliefs) and moral (rather than epistemic) responsibility, but it seems clear that his answer would be the same, that is, he would reject (z). We can also examine the set {x'yc}:

(x') Our actions are beyond voluntary control;

(y) Our actions are subject to normative censure;

(c) 'Ought implies Voluntary-Can'.

Of course, not all actions are beyond voluntary control, and so we have to restrict our scope to the ones that are to derive an inconsistent set parallel to {abc}. However, as discussed in §3.2, Hume thinks that involuntary actions are only subject to normative censure under certain conditions (i.e. when they accurately reveal underlying character). It seems more systematic to say that Hume denies (c) for all cases of involuntary actions but nevertheless thinks

 © Qu 2016

Hume recommends analogous solutions to both puzzles: he rejects (c) and (z) respectively. Regarding doxastic involuntarism, I argued that Hume rejects 'Ought implies Voluntary-Can'—he recognises that we can and do hold involuntary beliefs normatively accountable. The {x, y, z} parallel to this would be to read Hume as rejecting 'Ought implies Libertarian-Can': according to Hume, we can hold deterministically caused actions normatively accountable. And clearly Hume does take this route:

> A man who is robbed of a considerable sum; does he find his vexation for the loss any wise diminished by these sublime reflections? Why then should his moral resentment against the crime be supposed incompatible with them? Or why should not the acknowledgment of a real distinction between vice and virtue be reconcileable to all speculative systems of philosophy, as well as that of a real distinction between personal beauty and deformity? Both these distinctions are founded in the natural sentiments of the human mind: And these sentiments are not to be controuled or altered by any philosophical theory or speculation whatsoever. (EHU 8.35; SBN 102-3)

Here Hume makes the point that we do in fact hold deterministic actions normatively accountable, regardless of any 'sublime reflections' or philosophical principles (such as 'Ought implies Libertarian-Can') that say otherwise. Compare this passage with what Hume says regarding the involuntariness of moral character:

> But this, in the mean time, must be allowed, that sentiments are every day experienced of blame and praise, which have objects beyond the dominion of the will or choice, and of which it behoves us, if not as moralists, as speculative philosophers at least, to give some satisfactory theory and explication. (EPM App 4.21; SBN 322)

Similarly, here Hume points out that we do hold involuntary actions normatively accountable, regardless of what 'philosophers, or rather divines under that disguise' say to the contrary; what matter to moral evaluation are simply our sentiments of approbation and blame. Although his accounts of normativity with respect to determinism and involuntariness differ in some respects, it is clear that we can draw striking parallels between the two.

A brief point to note: perhaps tellingly, Hume describes his project here as that of a 'speculative philosopher'—even when dealing with the prescriptive notion of normativity, Hume stresses his descriptive

---

that some involuntary actions are morally irrelevant for external reasons due to his virtue-based understanding of morality.

 © Qu 2016

project of providing a science of man. And given that we do in fact pass normative judgments not only on deterministic actions but also involuntary beliefs, Hume's account is committed at the very least to allowing for their being held normatively accountable.

5. Conclusions and implications

I will end the paper by exploring a few implications of the account contained within. First, what are the implications of this for Hume and 'Ought implies Can'? I will note that the arguments of this paper do not preclude Hume from holding an even weaker version of 'Ought implies Can'. Hume seems to hold 'Ought implies Can' in a very weak sense of 'can', and only limited to actions, that is, 'Ought implies Can' in the sense that the fact that we ought to perform an action implies that human nature in general contains a motive or passion capable of generating this action:

> No action can be requir'd of us as our duty, unless there be implanted in human nature some actuating passion or motive, capable of producing the action. (THN 3.2.5.6; SBN 518)

Needless to say, this is perfectly compatible with holding involuntary, deterministic beliefs normatively accountable, since this principle applies only to actions.

Perhaps a more interesting implication has to do with the relationship between moral and epistemic normativity for Hume. Unlike the structure of Hume's account of morality, the structure of Hume's epistemic account is less well understood. I have drawn close parallels between the two normative structures, which suggest the possibility that beliefs are evaluable on the basis of the epistemic character traits that they exemplify; I explore such an account in more detail in my 'Hume's Practically Epistemic Conclusion?' (2014a). More generally, these close parallels between the two might lend some plausibility to 'practical' interpretations of Hume's epistemology, which take epistemic justification to be grounded in moral justification. Commentators such as Owen (1999) and McCormick (2005) who endorse both 'Prior Voluntary Action' and practical interpretations (in the same works, no less!) might find that this paper provides reasons to favour my account beyond the arguments I provided; if one takes seriously the parallel (or even identity) between moral and epistemic normativity for Hume, it seems unclear why we should think 'Ought implies Can' applies to one but not the other. Elsewhere in 'Hume's Practically Epistemic Conclusions?' (2014a), I argue that practical interpretations prove problematic as an

 © Qu 2016

interpretation of THN 1.4.7 given Hume's commitments elsewhere in the *Treatise*, but I have recently come to think that given the mess Hume got himself into at this point in the *Treatise*, it might be plausible that he simply propounds a problematic position here. There is much more to be said on this issue, but this paper is not the place to do so.[44]

References

Alston, William 1988: 'The Deontological Conception of Epistemic Justification'. *Philosophical Perspectives* 2.

Audi, Robert 2001: 'Doxastic Voluntarism and the Ethics of Belief'. In Steup 2001.

——2008: 'The Ethics of Belief: Doxastic Self-control and Intellectual Virtue'. *Synthese* 161.

Beebee, Helen 2006: *Hume on Causation*. Routledge.

Chuard, Philippe, and Nicholas Southwood, 2009: 'Epistemic Norms without Voluntary Control'. *Nous* 43:4.

Falkenstein, Lorne 1997: 'Naturalism, Normativity, and Scepticism in Hume's Account of Belief'. *Hume Studies* 23:1.

Feldman, Richard 1988: 'Epistemic Obligations'. *Philosophical Perspectives* 2.

——2000: 'The Ethics of Belief'. *Philosophy and Phenomenological Research* 60:3.

——2001: 'Voluntary Belief and Epistemic Evaluation'. In Steup 2001.

Flew, Antony 1961: *Hume's Philosophy of Belief*. Routledge and Kegan Paul Ltd.

Garrett, Don 1997: *Cognition and Commitment in Hume's Philosophy*. Oxford/New York: Oxford University Press.

Hahn, Lewis (ed.) 1997: *The Philosophy of Roderick Chisholm*. Lasalle, Ill.: Open Court.

[44] This paper originated as a chapter in my B.Phil. thesis, and benefitted immensely from uncountably many comments from my B.Phil. supervisor Peter Millican, to whom an overwhelming debt is owed. The paper was subsequently developed and went on to form a part of my Ph.D. thesis, and consequently I am incredibly grateful to my Ph.D. supervisor Don Garrett for the significant role he played in this stage of the paper's development; much thanks also to Béatrice Longuenesse, Jim Pryor and Robert Hopkins, who provided detailed and helpful comments. I also wish to express my heartfelt thanks to Ryan Hickerson, whose insightful comments and stimulating correspondence greatly improved the paper. I am also grateful to an audience at the National University of Singapore for excellent questions and discussion. Much thanks also to four anonymous referees for *Mind*, as well as the journal editor, for excellent and incisive comments.

 © Qu 2016

Hickerson, Ryan 2013: 'What the Wise Ought Believe: a Voluntarist Interpretation of Hume's General Rules'. *British Journal for the History of Philosophy* 21:6.

Hume, David 1739-1740: *A Treatise of Human Nature.* L. A. Selby-Bigge (ed.), 2nd ed. revised by P.H. Nidditch. Oxford: Clarendon Press, 1975.

——1748: *Enquiries concerning Human Understanding and concerning the Principles of Morals.* L. A. Selby-Bigge (ed.), 3rd edition revised by P. H. Nidditch. Oxford: Clarendon Press, 1975.

Huss, Brian 2009: 'Three challenges (and Three Replies) to the Ethics of Belief'. *Synthese.*

Kemp-Smith, Norman 1941: *The Philosophy of David Hume.* Macmillan and Co., Limited.

Leon, Mark 2002: 'Responsible Believers'. *The Monist* 85.

Loeb, Louis E. 1977: 'Hume's Moral Sentiments and the Structure of the *Treatise*'. *Journal of the History of Philosophy* 15.

——2002: *Stability and Justification in Hume's* Treatise, Oxford University Press.

MacIntosh, J.J., and H.A. Meynell, (eds) 1994: *Faith, Scepticism and Personal Identity.* University of Calgary Press.

Marusic, Jennifer S. 2010: 'Does Hume Hold a Dispositional Account of Belief?'. *Canadian Journal of Philosophy* 40:2.

McCormick, Miriam 2005: 'Why Should We Be Wise?'. *Hume Studies* 31:1.

Millican, Peter (ed.) 2002a: *Reading Hume on Human Understanding: Essays on the First Enquiry.* Oxford: Clarendon Press.

Millican, Peter (ed.) 2002b: *'Hume's Sceptical Doubts concerning Induction'.* In Millican 2002a.

Morris, William Edward 2000: 'Hume's Conclusion'. *Philosophical Studies* 99:1.

Norton, David 1982: *David Hume.* Princeton University Press.

——1994: *'How a Sceptic May Live Scepticism'.* In J.J. MacIntosh, and H.A. Meynell 1994.

——2002: 'Of the Academical or Sceptical Philosophy'. In Millican 2002a.

Owen, David 1999: *Hume's Reason.* Oxford: Oxford University Press.

Owens, David 2000: *Reason Without Freedom: The Problem of Epistemic Normativity.* Routledge.

Passmore, John 1980: *Hume's Intentions.* 3rd edition, Gerald Duckworth & Co. Ltd.

Price, Henry 1969: *Belief.* George Allen & Unwin Ltd.

 © Qu 2016

Qu, Hsueh 2014a: 'Hume's Practically Epistemic Conclusions?'. *Philosophical Studies* 170:3, pp. 501–24.
—— 2014b: 'Hume's Positive Argument on Induction'. *Nous* 48:4, pp. 595–625.
—— forthcoming: 'Hume's Internalism in EHU 12'. Forthcoming in *Philosophy and Phenomenological Research*.
Reid, Thomas 1785: *Essays on the Intellectual Powers of Man*. Derek R. Brookes (ed.). Edinburgh University Press, 2002.
Rescher, Nicholas 1993: 'Moral Luck'. In Statman 1993.
Richards, Nicholas 1986: 'Luck and Desert'. *Mind*, 65: 198–209.
Ridge, Michael 2003: 'Epistemology Moralized: David Hume's Practical Epistemology'. *Hume Studies* 29:2.
Ryan, Sharon 2003: 'Doxastic Compatibilism and the Ethics of Belief'. *Philosophical Studies* 114.
Saka, Paul 2000: '"Ought" does not imply "Can"'. *American Philosophical Quarterly*, 37.
Sinnott-Armstrong, Walter 1984: '"Ought" conversationally implies "Can"'. *The Philosophical Review* 93.
Statman, D. (ed.) 1993: *Moral Luck*. Albany: State University of New York Press.
Steup, Matthew (ed.) 2001: *Knowledge, Truth, and Duty*. Oxford University Press.
Stocker, Michael 1971: '"Ought" and "Can"'. *Australasian Journal of Philosophy* 39.
Thomson, Judith Jarvis 1993: 'Morality and Bad Luck'. In Statman 1993.
Vitz, Rico 2009: 'Doxastic Virtues in Hume's Epistemology'. *Hume Studies* 35.
Williams, Bernard 1981: *Moral Luck*. Cambridge: Cambridge University Press.
Williams, Michael 2004: 'The Unity of Hume's Philosophical Project'. *Hume Studies* 30:2.
Wilson, Fred 1997: *Hume's Defence of Causal Inference*, University of Toronto Press.
Wolterstorff, Nicholas 1997: 'Obligations of Belief: Two Concepts'. In Hahn 1997.

 © Qu 2016

Cognitive Mobile Homes

Daniel Greco
Yale University
daniel.greco@yale.edu

While recent discussions of contextualism have mostly focused on other issues, some influential early statements of the view emphasized the possibility of its providing an alternative to both coherentism and traditional versions of foundationalism. In this essay, I will pick up on this strand of contextualist thought, and argue that contextualist versions of foundationalism promise to solve some problems that their non-contextualist cousins cannot. In particular, I will argue that adopting contextualist versions of foundationalism can let us reconcile Bayesian accounts of belief updating with a version of the holist claim that all beliefs are defeasible.

1. Introduction

Epistemic contextualists often motivate their position by arguing that contextualism provides a satisfying resolution of certain sceptical paradoxes.[1] In this essay, however, I will present one version of a very different strategy for motivating contextualism.[2] The strategy I will explore attempts to motivate contextualism by arguing that it provides an appealing way to chart a middle course between coherentism on the one hand, and traditional, non-contextualist versions of foundationalism on the other[3].

[1] For paradigm instances of this strategy, see Stine (1976), Cohen (1987), DeRose 1995) and Lewis (1996).

[2] Actually, as will become clear later, while I will focus on contextualism for ease of exposition, my hope is that the strategy can be employed by non-contextualists of various stripes as well. Epistemic relativists (MacFarlane 2011a, 2011b) and expressivists (Chrisman 2007) have all offered diagnoses of various phenomena that are structurally similar to contextualist diagnoses. My hope is that any view that can mimic familiar contextualist treatments of, for example, sceptical paradoxes (as the relativist can, by appealing to assessment-sensitivity in many of the places where the contextualist would appeal to context-sensitivity) can also mimic the contextualist treatment of foundationalism I will offer here, and that the issues that separate contextualism from a variety of rival views will, for the most part, not be relevant to the present discussion.

[3] See Williams (1977) and Annis (1978). Arguably, it was also anticipated by Wittgenstein (1969). In his recent book-length defence of contextualism, DeRose (2009, pp. 21–2) brings up

Mind, Vol. 126 . 501 . January 2017
doi:10.1093/mind/fzv190 Advance Access publication 29 October 2016
© Greco 2016

These two strategies are independent of one another—one can hold that contextualism provides a satisfying resolution of sceptical paradoxes without holding that it provides any special advantage to the foundationalist,[4] just as one can endorse contextualist versions of foundationalism without using them to offer the familiar contextualist treatment of sceptical paradoxes.[5] I focus on the less common strategy because I believe it points the way to some surprising further potential applications of contextualism that do not suggest themselves quite as readily when contextualism is motivated via sceptical paradoxes.

2. Foundationalism

I will use the term 'foundationalism' to refer to a view about the structure of epistemic justification. In particular, I will use it to refer to the view that both of the following claims hold: (1) at least some beliefs are justified even without receiving support from other beliefs—I will call such beliefs, should any exist, 'foundational'—and (2) all justified beliefs are either foundational or derive all their support (perhaps indirectly) from justified foundational beliefs.[6]

A natural question for the foundationalist is the following: what criteria must beliefs meet in order to be foundational? Historically influential versions of foundationalism held that beliefs must meet very strict criteria in order to be foundational—perhaps they must be impossible to doubt, or must have contents that are necessarily true whenever believed, or must have some other strong modal properties. It is easy to see why one might find such properties epistemologically interesting. If it were possible to rationally reconstruct our entire body of beliefs from some sparse set of claims that are impossible to doubt, or that must be true (given that we believe them), doing so would be a

---

the idea that contextualism has some special relevance to the debate between foundationalism and coherentism, only to set it aside. I agree with DeRose that there is no entailment from contextualism to either foundationalism or coherentism. But as the rest of this essay should make clear, I do think that contextualism and foundationalism complement each other nicely.

[4] As DeRose does in the passage I refer to in footnote 1.

[5] For instance, while Michael Williams (2013) does endorse a sort of contextualist version of foundationalism—he calls it 'formally foundationalist'—he offers a very different treatment of scepticism from that found in the writings of Stine, Cohen, DeRose and Lewis. In particular, he does not concede that there is *any* context in which global sceptical challenges are successful.

[6] While 'foundationalism' is used in too many different ways to speak of a 'standard' use, my formulation is similar to what many writers call 'minimal' foundationalism (Alston 1976).

 © Greco 2016

great intellectual achievement. Criticisms of traditional versions of foundationalism have tended not to target the desirability of completing such a project, but the feasibility. It is not clear whether *any* beliefs have such strong modal properties, and even if some do (such as, perhaps, the conclusion of Descartes's Cogito), they are almost certainly too few to serve as an adequate foundation for the rest of what we normally take ourselves to be justified in believing.

In light of these considerations, contemporary foundationalists tend to defend views on which the requirements for foundational beliefs are more lax, and in particular, need not be free from the threat of doubt.[7] In this first section, however, I want to show how one might be motivated to retreat, not to a laxer invariantist form of foundationalism, but instead to a view according to which the criteria beliefs must meet in order to be foundational vary with context.

Suppose we are sympathetic to the Cartesian idea that freedom from doubt really is the *sine qua non* of foundational status. After all, we will not be happy treating the claim that *P* as though it is foundational—for example, justifying other beliefs by showing how they are supported by the claim that *P*—if we doubt that *P*. If we allow that which beliefs play the foundational role may vary from context to context, however, then we can insist that foundations must be undoubt*ed*, without taking the Cartesian position that they must be undoubt*able*. We might summarize the two ways of respecting the idea that foundations must be free from doubt as follows:

> *Cartesian*: For a belief to be foundational, it must be undoubt*able*.
>
> *Contextualist*: For a belief to be foundational (in a context), it must be undoubt*ed* (in that context).

Both the Cartesian and the contextualist agree that our justified beliefs rest on a foundation of undoubted premisses. The Cartesian thinks there are certain types of beliefs that can never be in doubt, and so can always play a foundational role. The contextualist disagrees—because she thinks all beliefs are dubitable, she does not think that any beliefs can play a foundational role in *every* context.[8] So while she will agree with the Cartesian in insisting that foundations be undoubted, she

[7] For example, Gilbert Harman (2001) defends a view on which *all* beliefs count as foundational. Phenomenal conservatives defend the only slightly less permissive view that all beliefs that *seem* true to their subjects are foundational for those subjects.

[8] Or if she grants that *some* beliefs are indubitable, she thinks they are too few to serve the Cartesian's purposes.

 © Greco 2016

only requires *contextually local* freedom from doubt, rather than immunity to doubt across all contexts.

A natural thought for the contextualist foundationalist is that when a belief counts as 'foundational' in a context, it is treated by the occupants of that context as (at least provisionally) a default assumption in reasoning and argument. Foundationalism is plausible because we must always treat *some* beliefs as default assumptions in reasoning and argument. But it is also plausible that *no* belief will (or should) be treated as a default assumption for reasoning and argument in *every* context. Rather, any belief that we treat as a default assumption in one context we can treat as a mere hypothesis in another, to be accepted only if it can be supported on independent grounds—that is, only if it receives support from what we treat as default assumptions in our *new* context.

While the metaphor of Neurath's raft is often appealed to by coherentists, I think it is ultimately most congenial to the contextualist foundationalist. The metaphorical way of putting the point of the previous paragraph is as follows: at any given time, we must stand *somewhere* on the raft, and we cannot examine the planks on which we are currently standing: each context treats *some* beliefs as foundational, so the coherentist is wrong. But still, we can always shift our weight to new planks, so as to be able to examine the ones on which we were previously standing: no belief counts as foundational in every context, so the non-contextualist foundationalist is wrong.

Why think of contextualist foundationalism as a sort of middle course between traditional versions of foundationalism on the one hand, and coherentism on the other? While others have defended this claim at greater length—again, see Williams (1977) and Annis (1978)—the rough idea is that contextualist foundationalism is meant to avoid the most serious pitfalls for the two more traditional views about the structure of justification. With the traditional foundationalist, the contextualist foundationalist can accept the attractive claim that justification by coherence alone is impossible, because it would be circular. But with the coherentist, the contextualist foundationalist can agree that there is no distinctive class of beliefs well-suited to play the regress-stopping role required by foundationalism.[9] Ultimately, however, my aim in this paper will not be to argue for

[9]The emphasis is on 'distinctive' here. The contextualist will accept that, in any given context, there are regress stoppers. But the regress stoppers in one context may have little in common with the regress stoppers in another. There will not be some distinctive property—such as indubitability—shared by all regress stoppers in all contexts.

 © Greco 2016

contextualist foundationalism's superiority to coherentism and traditional foundationalism, but to explore an application of it in the context of formal epistemology.

There are many ways of developing the idea of contextualist foundationalism, not all of which would be labelled 'contextualism' in contemporary parlance. The following three views bear a family resemblance to one another, and each, in my opinion, can be motivated by the considerations that I discuss in this essay. While I will generally focus my discussion on what I will call 'orthodox contextualist foundationalism', this is for ease of exposition; most of what I say should be applicable to each of the views I am about to mention, *mutatis mutandis*.

> *Orthodox Contextualist Foundationalism:*
> Sentences of the form '*S*'s belief that *P* is justified' (as well as sentences attributing foundational justification more narrowly) express different propositions when uttered in different conversational contexts. Features of conversational contexts that affect which propositions such sentences express may include the presuppositions made by the participants in a conversation, the purposes of the conversation, and the practical situation faced by the participants in the conversation. In general, a proposition *P*'s being presupposed by the parties in a conversational context *C* tends to make sentences of the form '*S*'s belief that *P* is foundational' true, when uttered in *C*.[10]

> *Relativist Foundationalism*:
> Sentences of the form '*S*'s belief that *P* is justified' (as well as sentences attributing foundational justification more narrowly) express propositions that are not absolutely true, but true only relative to contexts of assessment. Which contexts of assessment we take on for the purpose of evaluating sentences about foundationality (and which sentences about foundationality we will be prepared to endorse) may depend on factors such as which presuppositions we and our interlocutors make, our purposes in evaluating the attribution, and our practical situation (along with, perhaps, the practical situation of the subject of the evaluation).[11]

[10] See Stine (1976), Cohen (1987), DeRose (1995) and Lewis (1996). For the emphasis on presupposition in particular, see Blome-Tillmann (2009).

[11] I intend this to be in the spirit of MacFarlane (2011b).

 © Greco 2016

Expressivist Foundationalism:
Sentences of the form '*S*'s belief that *P* is justified' (as well as sentences attributing foundational justification more narrowly) do not express propositions of the usual sort (except perhaps in a deflationist sense), but are instead used to express our acceptance of epistemic norms. As a matter of anthropological fact, which epistemic norms we accept (and so which justification attributing sentences we will be prepared to endorse) varies with factors such as which presuppositions we and our interlocutors make, our purposes in evaluating attributions of foundationality, and our practical situation.[12,13]

I have characterized the above views as views concerning the truth conditions of sentences of the form '*S*'s belief that *P* is justified' or '*S*'s belief that *P* is foundationally justified'. While this might seem like a simple extension of familiar views about sentences of the form '*S* knows that *P*', in fact things are not so simple; unlike 'knows', 'justified' and 'foundational', at least in the context of epistemology, are arguably technical terms.[14] While it is easy to see how one might appeal to empirical linguistic data to adjudicate on debates about the semantics of 'knows', it is harder to see how similar debates can even get off the ground concerning terms that do not occur in ordinary usage. My hope is that while I have expressed the views above as views about the truth conditions of a certain sort of sentence, it is not so hard to see how we can understand them as views about what it takes for beliefs to have certain sorts of epistemic statuses—statuses that *could* be attributed by sentences containing words like 'justified' and 'foundational', but which can play important theoretical roles even if we do not often directly attribute them. For instance, if I take some belief of yours to have one of the relevant statuses (e.g. to be justified), that will likely have implications for how I interact with you—perhaps I will be willing to take the belief for granted in the course of planning our joint endeavours, and will not object when you express the belief in assertion, or assert consequences of the belief—even if I never directly attribute that status. And even if we rarely

[12] I intend this to be in the spirit of Chrisman (2007).

[13] It may also be the case that epistemic contrastivists, such as Schaffer (2004, 2007), can take advantage of some of the arguments I will discuss.

[14] Cohen (1995) argues that 'epistemic justification' is a technical term, in ways that contrast with 'knowledge'.

 © Greco 2016

explicitly attribute foundationality (or its lack) to beliefs, we may manifest judgements about which beliefs are foundational in other ways. For instance, if I assert, 'Bob must be in his office', it is plausible that I am not only expressing my belief that Bob is in his office, but also conveying that I do *not* take this belief of mine to be foundationally justified.[15]

There are many important differences (and perhaps some not so important ones as well) between orthodox contextualism, relativism and expressivism, and these differences have been much debated outside the context of their relevance to foundationalism.[16] While the debates have usually concerned knowledge attributions rather than attributions of foundationality, the issues are similar. In fact, given certain additional premisses, contextualism, relativism and expressivism about knowledge *just are* versions of contextualism, relativism and expressivism about foundations. For instance, Williamson (2000) is naturally interpreted as a foundationalist who holds that a belief that *P* is foundational for a subject *S* just in case *S* knows that *P*. If we accept the view that knowledge plays this foundational role, then contextualism about knowledge is contextualism about foundations. While Williamson (2005) has himself argued against contextualism, the suggestion that Williamsonian views about the centrality of knowledge in epistemology can be fruitfully combined with contextualism is not unheard of (see Ichikawa 2013).

My strategy in this paper, then, will be to try to remain as neutral as possible on the issues that separate different versions of 'shifty' epistemological views. My hope and expectation is that these differences will not substantially affect how, as a class, these views interact with foundationalism. In particular, I expect that each of these views promises to offer some similar advantages over traditional versions of foundationalism, whatever other drawbacks they may have. For this reason, while I will focus on arguing that orthodox contextualist

[15] I am gesturing towards the view that 'must' in English is an 'evidential', in that (one of) the roles of 'must' in sentences of the form 'It must be that *P*' is for the speaker to indicate the nature of her evidence for her belief that *P*—in particular, to indicate that it is somehow indirect. While the view that 'must' is an evidential is common, it is controversial just what about a speaker's evidence she conveys by using 'must'. See Aikhenvald (2004) for a general discussion of the category of evidentials.

[16] Chrisman (2007) argues that expressivism can reap the benefits of contextualism without some of the main costs (though see Plunkett and Sundell 2013 for a response to this style of argument). Field (2009) expresses sympathy for both expressivism and relativism in epistemology, and is doubtful that, properly understood, the views ultimately conflict with one another.

 © Greco 2016

foundationalism can address certain problems that non-contextualist versions of foundationalism cannot, I expect that similar arguments could be constructed for the other two views listed above.

So far, I have briefly motivated the idea that some version of contextualist foundationalism is a natural response to the infeasibility of the Cartesian project. In the next section of this essay I will argue that contextualist foundationalism promises to bear some surprising fruit; in particular, I will argue that it can resolve a persistent puzzle in Bayesian confirmation theory.

Before jumping into the details, however, it may help to start with a bird's eye view of the terrain. There is a certain sort of traditional foundationalism that fits very naturally with Bayesian views about rational belief updating. Very roughly, it goes like this. There is a domain to which we have a sort of privileged epistemological access—at least, if we're paying attention, we always have it in our power to know the truth concerning matters in this domain. It constitutes, as Williamson (2000, p. 92) puts it, our 'cognitive home', where nothing is hidden to view. Because we clearly *can* be mistaken about almost everything, even when doing our best to be careful, this domain will have to concern our own minds. According to this view, while we *can* be mistaken about whether we are injured or whether we are seeing a red apple, we cannot be mistaken (at least if we're paying attention) about whether we are in pain or whether we are having a visual experience as of seeing a red apple. We can think of our belief-forming processes as a journey: we start in our cognitive home (where we can know the truth) and then we venture outside, forming conjectures about further matters based on how things are in our home. For example, if things are a certain way in our home (we have an experience as of a red apple in front of us), then they are *probably* a certain way outside our home (there probably *is* a red apple).[17]

[17] The project of providing an inference to the best explanation-style response to scepticism, in which sceptical hypotheses are rebutted by arguing that commonsense hypotheses provide better explanations of some class of mental phenomena—pursued by, among others, BonJour (1985) and Vogel (1990)—seems to me easiest to motivate in the context of the above traditional foundationalist picture. The question of whether, or to what degree, commonsense hypotheses provide a better explanation of facts about some class of mental phenomena than sceptical hypotheses is most obviously of epistemological import if we are straightforwardly in a position to know facts about the mental phenomena, and are justified in accepting further hypotheses only to the degree that they are supported by those mental facts (for example, only to the degree that they can provide good explanations of those facts). Once this assumption is abandoned, it is less clear what the epistemological significance would be of the success or failure of the explanationist project. While BonJour (1985) does explicitly defend coherentism and argue against foundationalism, it seems to me that the difficulties he

 © Greco 2016

This version of the foundationalism view fits neatly with Bayesian accounts of belief updating, because together they suggest a natural answer to the question of how a rational agent should update her beliefs over time: she should strictly conditionalize on propositions concerning matters in her cognitive home. That is, Bayesianism provides a straightforward way of formalizing the idea that hypotheses concerning matters outside our home are credible only to the degree that they are supported by facts about our home, namely, one's credence in any proposition should equal the conditional probability of that proposition, given all the facts about how things are in one's home.[18]

So the view that we have cognitive homes and the view that strict conditionalization is the appropriate way for a rational agent to update her beliefs together form a neat package. But what if we lack cognitive homes? What if there is no domain in which careful reflection is guaranteed to lead to the truth? In the next section, I will argue that the neat fit between Bayesianism and foundationalism does not require that we have cognitive homes of the traditional sort. There need not be a domain about which we *cannot* be wrong, so much as, in each context, a domain about which it is taken for granted that we're *not* wrong. We can think of this domain as a *kind* of cognitive home, so long as we recognize the differences between the cognitive homes we'll be positing and the more familiar ones. In some contexts, a great deal might count as part of our home (including facts concerning matters beyond our own minds), while in others, even facts about whether or not we are in pain might fail to make it in. We have cognitive homes, but because they shift from context to context, we can think of them as mobile homes. These cognitive mobile homes are, I will argue, sufficient to save the package of Bayesianism and foundationalism.

3. Bayesianism and defeat

A number of writers have argued that there is a tension between Bayesian accounts of belief updating and the holist epistemological claim that all beliefs are defeasible (see Christensen 1992, Weisberg

---

encountered in that work in justifying the 'doxastic presumption', as well as the fact that he later reversed his position and endorsed foundationalism (BonJour 1999)—while still defending inference to the best explanation as a response to scepticism—reinforce the idea that explanationist responses to scepticism fit most neatly with foundationalism.

[18] Thanks to an anonymous referee for help on how to put this point.

 © Greco 2016

2009 and Pryor 2013). In this section, I will first introduce the topic of defeat, and will try to show why, prima facie, Bayesian approaches to belief updating promise to deliver an attractive account of how defeat works. I will then introduce an apparent limitation of such approaches: if we grant the holist claim that all beliefs are defeasible, then there will be cases of defeat that extant Bayesian approaches seem to lack the resources to model. At best, this looks like a gap waiting to be filled with some new extension of the Bayesian framework. At worst, it looks like evidence that the whole approach is misguided. I will argue that we can solve the problem without any novel formal machinery once we view Bayesian theories of belief updating as species of foundationalism and recognize that this leads to the possibility of a contextualist Bayesian foundationalism. The argument that there *is* a difficult problem for the Bayesian depends, I will argue, on the assumption that the Bayesian is a non-contextualist foundationalist. If the Bayesian is a contextualist foundationalist, however, the problem has a very natural solution.

3.1 Defeat

Suppose you read in a newspaper that the majority of Salemites intend to vote 'Yes' on Proposition 17. As a result, you come to believe that Proposition 17 will pass. Later, however, you read that the 'No' movement is about to release an advertising campaign that, based on testing in focus groups, looks to be extremely effective at changing people's minds.

After you read the first story but before you read the second, you have justification to believe that Proposition 17 will pass. But after you read the second story, you no longer have such justification. Epistemologists say that reading the second story 'defeats' the support that you had for the claim that Proposition 17 will pass.

Can we say anything more precise about defeat? Suppose E is our initial evidence, H is the hypothesis that evidence supports, and E' is our defeater. A natural thought is that in cases of defeat, while E on its own is good evidence for H, $E \wedge E'$ is not.

Moreover, the probability calculus suggests a natural formalization of these relations. After all, probabilistic confirmation is non-monotonic; E can confirm H, even though $E \wedge E'$ does *not* confirm H.[19]

[19] This is not intended to be a *requirement* for defeat—for example, E' might defeat E's support for H without doing so *fully*, so $E \wedge E'$ might still provide some confirmation for H, even though E' is a defeater for E's support for H. Thanks to an anonymous referee for pointing this out.

 © Greco 2016

Because of the non-monotonicity of confirmation, we can treat cases of defeat as ones in which an initial body of evidence supports some hypothesis, but a larger body of evidence—one that includes a defeater in addition to the initial evidence—does not. Applied to the example just presented, the strategy might go as follows.

> *Intention* = The majority of Salemites intend to vote 'Yes' on Proposition 17.
>
> *Pass* = Proposition 17 will pass.
>
> *Ad Campaign* = A persuasive ad campaign attacking the proposition is about to be released.

Intention is evidence for *Pass*. We can represent this probabilistically by saying that $P(Pass \mid Intention) > P(Pass)$. The second story is a defeater for the claim that the proposition will pass. If we assume that the defeater brings us back to roughly the level of support we had initially for *Pass*, we can represent this as follows: $P(Pass \mid Intention \wedge Ad\ Campaign) \approx P(Pass)$.

While this is a promising beginning of a Bayesian account of defeat, a full story would have to say much more.[20] The challenge to Bayesian accounts of defeat that I will discuss, however, challenges any probabilistic account of defeat that turns on the non-monotonicity of probabilistic confirmation, as the proto-account I have sketched so far does.

3.2 The ubiquity of defeasibility

The (putative) difficulty for Bayesian accounts of defeat is the following: just as one can undermine support for a hypothesis by attacking the link between one's evidence and the hypothesis, one can also undermine support for a hypothesis by attacking the evidence itself. So not all defeat can be understood as involving the accumulation of evidence, where the final, total body of evidence fails to support some hypothesis that was supported by the initial, smaller body of evidence. Arguments of this general form have been given by David Christensen (1992), Jonathan Weisberg (2009) and James Pryor (2013). Exactly what this amounts to, and why it has been thought to be a problem for the Bayesian, can be brought out with an elaboration of our earlier example.

[20] Kotzen (MS) does an admirably thorough job of displaying the resources the Bayesian has in describing a wide variety of cases of defeat.

 © Greco 2016

As before, you read that the majority of Salemites intend to vote 'Yes'. But now, the second story does not concern an ad campaign, but instead alleges that the polls on which the first article was based come from a source with a flawed methodology—their results are not likely to be representative of what the people of Salem, on the whole, intend.

Intuitively, this defeats your support for the hypothesis that Proposition 17 will pass. Why does this present a prima facie problem for the Bayesian? Let *Intention* and *Pass* have the same meanings as before, and let *Flawed Source* be understood as follows:

> *Flawed Source* = A second story reported that the polls relied on by the first story come from a source with a flawed methodology.

What we'd *like* to say would be that the following two conditions hold:

(1) $P(Pass \mid Intention) > P(Pass)$

(2) $P(Pass \mid Intention \wedge Flawed\ Source) = P(Pass)$

This would be to treat *Flawed Source* as playing the same role that *Ad Campaign* did in our earlier case. The problem is that the second condition *does not* hold. If the majority of Salemites intend to vote 'Yes', then regardless of how the polls were conducted, the proposition will probably pass. That is, $P(Pass \mid Intention \wedge Flawed\ Source) > P(Pass)$.

The basic problem is that if we treat the evidence provided by the first story as *Intention*, then that evidence is not defeated by the story about the flawed polling methodology. Rather, if we are to accommodate the possibility of defeat via that story, it looks as though we need to treat the evidence provided by the first story as something like *Polls*:

> *Polls* = A poll reported that the majority of Salemites intend to vote 'Yes'.

Only if something like *Polls* was the evidence that the first story provided can we explain why the story about the flawed methodology defeats your support for thinking that the proposition will pass.[21]

[21] One might think that we could stick with treating your evidence as *Intention* rather than as *Polls* so long as we treat you as Jeffrey conditionalizing on the partition [*Intention*, ~*Intention*], rather than strictly conditionalizing on *Intention*. Christensen (1992) and Weisberg (2009) convincingly argue that this will not work. While discussing the details is beyond the scope of this paper, the basic problem is as follows. In order to allow that *Flawed Source* undermines the support that *Intention* provides for *Pass*, *Flawed Source* and *Pass* must start out probabilistically independent of one another, but must become probabilistically dependent (in particular, they must become negatively relevant to one another) after Jeffrey conditionalizing on the partition [*Intention*, ~*Intention*]. But Jeffrey conditionalization cannot induce this sort of probabilistic dependence. More intuitively, conditionalizing

 © Greco 2016

So far, this might not seem like a problem at all. What's wrong with treating the evidence provided by the first story as *Polls*? The danger is that just as *Intention* can be undermined, so too can *Polls*. For instance, suppose the second story is just a correction—it reports that the first story misreported the poll results. If this defeater is to be accommodated, the evidence will need to concern, not what the polls said, but what the newspaper said the polls said.

Perhaps, at some point, we'll get to a way of characterizing the evidence provided by the first story such that it is immune to the sorts of defeat we've been discussing. Maybe the evidence will concern, say, the sorts of experiences I have upon reading it. But to many philosophers, especially those sympathetic to broadly holist epistemological positions, this will seem quaintly optimistic. Even claims about our experiences are not immune to reasonable doubt, and can be defeated in the right circumstances.[22] To hold out hope that there is some class of beliefs that is immune to rational undermining, and is such that all episodes of learning can be understood as involving the acquisition of some new beliefs in the privileged class (with further changes in our body of beliefs coming via conditionalization), is to hope for something very much like the traditional, 'cognitive home' version of foundationalism discussed at the end of the previous section. To be clear, there are certainly contemporary philosophers who

---

on *Intention*—whether strictly or Jeffrey-style—does not give us a way to represent that the support for *Intention* depends on beliefs about the polls, and so does not give us a way to represent that the support for *intention* can be defeated by information about polls. To be sure, one *can* deal with the problem if one allows oneself more flexibility in choosing an input partition—see Wagner (2013). For example, we might hold that when we read the newspaper, we do not simply Jeffrey conditionalize on [*Intention*, ~*Intention*], but perhaps instead use some more complex partition, including *Flawed Source*, *Pass*, and perhaps many other propositions. Weisberg is unhappy with this sort of response, however: 'Of course we can always set up the input partition to get the results we want, but then our input rule is doing all the work, and the inputs problem becomes the only interesting problem on the table' (2009, p. 805). Ultimately, I think both Wagner and Weisberg are on to something. Wagner is right that, given the right inputs to Jeffrey conditionalization (or, I'd point out, even strict conditionalization), these cases do not pose a problem for the Bayesian. Weisberg is right that, without some constraints on what count as appropriate inputs, this response only saves the Bayesian approach to updating at the cost of robbing it of explanatory power. Ultimately, I think this is where contextualism helps the Bayesian, in ways that should become clear in the latter part of this section.

[22] See Schwitzgebel (2011) for arguments to the effect that we are highly fallible introspectors. See also Williamson (2000), especially the discussion of our 'cognitive homelessness' in chapter 4, for more general arguments to the effect that there is no realm of facts which could play the foundational role that we might have hoped facts about experiences would play.

 © Greco 2016

hold such views—perhaps Smithies (2012a, 2012b) is the most explicit recent defender.[23] Moreover, there are projects taken seriously in philosophy that seem to me to presuppose such traditional foundationalist views.[24] I have no original arguments to offer in this paper against such views; still, I am interested in exploring the consequences for Bayesian accounts of belief updating once we grant their falsity.

To sum up, the challenge to Bayesian accounts of defeat is as follows. If the Bayesian is too generous about what our evidence is—for example, if she thinks that our evidence in the cases I have been discussing is *Intention*—then she will not be able to account for certain cases of defeat. But there's no way to avoid being too generous; whatever the Bayesian says our evidence is will turn out to be immune to certain intuitive instances of undermining according to the Bayesian. But nothing is immune to undermining.

Whether this is rightly thought of as a *challenge* at all will depend on what sort of ambitions we have for Bayesian accounts of defeat.[25] One response to examples like the ones I have been discussing is simply to concede that they are not cases to which extant Bayesian approaches are meant to apply. While this is fine as far as it goes, it raises the natural question of how we might supplement Bayesian approaches to account for *all* possible cases of defeat, rather than just some.[26] To be sure, there are substantial limitations of the Bayesian framework, and if we are to insist on a Bayesian approach to defeat, let alone one that is supposed to handle all possible cases of defeat, it should not be on

[23] For his response to the Schwitzgebel arguments mentioned in the previous footnote, see Smithies (2013).

[24] As mentioned in an earlier footnote, the project of responding to scepticism by appeal to inference to the best explanation seems to me to be one such example.

[25] Lydia McGrew (2013) argues that the seriousness of the problem I have been discussing 'depends entirely on what a given philosopher wanted from [Jeffrey conditionalization] in the first place'. While she grants that certain ambitious hopes are frustrated by the issues we've been discussing, she argues that this need not rob Jeffrey conditionalization of its interest: 'If, on the other hand, one regarded the [Jeffrey conditionalization] formula only as a useful epistemic tool for particular propositional updates, then the fact that a great many salient and interesting propositions can be updated using an obvious partition and Jeffrey's rule will be quite sufficient' (McGrew 2013, p. 581).

[26] Christensen (1992) seems hopeful that some extension of Bayesianism will prove adequate. Pryor (2013) considers some strategies for handling defeat in a version of the Bayesian framework, but also takes seriously the idea that significant departures from that framework might be required to account for defeat.

 © Greco 2016

the basis of the view that *all* phenomena of epistemological interest can be captured in the Bayesian framework.[27]

Still, there are reasons to be unhappy with the idea that an alternative formalism, or even an extension of the Bayesian approach we've been discussing so far, is what's called for here. First, alternative formalisms available so far run into similar troubles in accounting for undermining defeat (see Weisberg 2015). Second, the Bayesian account of defeat handles some cases—such as the first version of the newspaper case—extremely nicely, and these cases do not seem all that different from the cases where the Bayesian account seems to falter. If we rest content with the Bayesian treatment of the original case involving a story about polls and a story about an ad campaign, but look for some alternative approach to handle the variations on the original case, we risk ending up with an account of defeat that looks oddly disjunctive. At least to me, it seems that the sort of defeat provided by the story about the ad campaign is not all that different from the sort of defeat provided by the story about the flawed polling methodology, or the error in reporting the poll results. It would be nice to be able to treat each of these cases as having a similar formal structure, so if we're going to accept a Bayesian account of the first case (which is hard to resist), it would be nice to have a Bayesian account of the others as well. In the next section, I will argue that such an account can easily be had, so long as our Bayesianism takes a contextualist form.

3.3 Bayesianism as (contextualist) foundationalism

It is instructive to think of Bayesianism as a species of foundationalism, in which a subject's prior probability function, as well as her beliefs in the evidence propositions she updates on, both have a sort of foundational status. Just as non-Bayesian foundationalists hold that a subject's beliefs are justified just in case they derive support from the subject's foundational beliefs, Bayesians hold that a subject's credences are justified just in case they receive the right sort of support from the foundational elements in the Bayesian epistemological framework—that is, a subject's credences are justified just in case they result from conditionalizing the subject's prior probability function on the subject's evidence propositions.

[27] To take just one major example, since Bayesianism incorporates an assumption of logical omniscience, it cannot represent logical learning. While there have been attempts to model special cases of logical learning in a Bayesian framework (e.g. Garber 1983), I think it is fair to say that they are not generally regarded as likely to provide an attractive general treatment of logical learning in a Bayesian framework.

 © Greco 2016

One reason it is helpful to think of Bayesianism as a version of foundationalism is that it makes clear that the Bayesian faces versions of the familiar tricky question for non-contextualist foundationalists. Rather than the generic 'What criteria must beliefs meet in order to be foundational?' we can pose two Bayesian-specific versions of this question:

(1) What criteria must prior probability functions meet in order to be rationally permissible?

(2) What criteria must evidence propositions meet in order to be rationally updated upon?

The first question—referred to in the Bayesian literature as the 'problem of the priors'—is well known, though it has no generally agreed upon solution. Many historically popular answers to the first question, moreover, can be seen as paralleling historically popular answers to the generic question for foundationalists, and as sharing the difficulties of those answers. For example, Carnap's (1967) project of determining a prior probability function on purely logical grounds shares the appeal, and the infeasibility, of the traditional Cartesian foundationalist project.[28] And the orthodox subjective Bayesian view that *all* coherent prior probability functions are rationally permissible bears a family resemblance to Harman's view that all beliefs are foundational, and faces similar objections. From a bird's eye point of view, both views have the same structure: they say that rationality is a matter of reasoning appropriately *given* one's doxastic starting points. While they construe the starting points differently—the subjective Bayesian thinks of them as constituted by a subjective probability function, while Harman (1986) doubts that we have anything like degrees of belief in the sense countenanced by Bayesians—they both entail that what one ultimately ought to believe strongly depends on what one starts out believing, in ways that can seem implausible.[29]

[28] While there are many reasons to doubt that the Carnapian project can be completed, Nelson Goodman's (1983) objections are perhaps the best known.

[29] See Kelly and McGrath (2010). While they focus on objecting to the method of reflective equilibrium, they show how the considerations they raise apply to Harman's general foundationalism. Moreover, the objections they raise can be straightforwardly translated into objections to subjective Bayesianism. For example, somebody who starts out with a prior probability of 0 that humans will ever walk on the moon, and retains this degree of belief despite having access to the same evidence the rest of us have, seems *irrational*, even if their body of belief as a whole is probabilistically coherent.

 © Greco 2016

The second question—the question about which propositions subjects can rationally update on—is much less often discussed by Bayesians.[30] As the discussion of defeat above should make clear, however, it is not obviously any less difficult than the first; the arguments discussed above seem to suggest that *any* answer to the second question leads to the implausible conclusion that certain beliefs (namely, whichever ones we are entitled to update on) are immune to certain sorts of undermining.

Both of these questions for the Bayesian, however, can be given contextualist answers.[31] And while a contextualist treatment of the first question is a topic for another essay, in what follows I will argue that a contextualist treatment of the second question can let the Bayesian reconcile her treatment of defeat with a version of the ubiquity of defeasibility.[32]

[30] Deborah Mayo (1996, pp. 86–8) identifies the question of when one should accept a piece of evidence—in the Bayesian framework, when one should conditionalize on an evidence claim—as a crucial one for Bayesians, on which they often have little to say.

One notable exception is Timothy Williamson (2000). Like Mayo, he identifies the question as a key one for Bayesians. Rather than rejecting Bayesianism on the grounds that it cannot handle the question, however, he goes on to defend a version of Bayesianism on which a subject's evidential probabilities are determined by conditionalizing a prior probability function on each of the propositions the subject knows.

Still, to the extent that this view does *not* give a broadly Bayesian treatment of knowledge, it can be seen as a sort of retreat, at least if one starts from the point of view of an orthodox Bayesian philosopher of science. As Mayo notes, 'the need for a supplementary account of evidence would belie one of the main selling points of the Bayesian approach—that it provides a single, unified account of scientific inference' (1996, p. 88). However, if we're not antecedently committed to Bayesianism answering *all* questions in epistemology and philosophy of science (as I am certainly not), this sort of retreat need not strike us as unwelcome, even if we find the Bayesian framework a generally useful one.

[31] For some precedent, Jon Williamson (2010, p. 5) defends a version of Bayesianism on which 'an agent's evidence is relative to her current operating context', though ultimately Williamson's view has less in common with the sort of contextualism I am discussing than it might seem. In particular, he holds that whatever an agent takes for granted is part of her evidence—there's no room for us, as attributors of knowledge or justification, to hold that someone is treating something as part of her evidence when it is in fact not. The sort of contextualism I will be discussing does not have this commitment.

[32] While I know of no developed contextualist treatment of the problem of the priors, Timothy Williamson seems sympathetic to such a position in his treatment of evidential probability:

> The discussion will assume an initial probability distribution P. P does not represent actual or hypothetical credences. Rather, P measures something like the intrinsic plausibility of hypotheses prior to investigation; this notion of intrinsic plausibility can vary in extension between contexts. (Williamson 2000, p. 211)

Moreover, if Williamson adopted contextualism about knowledge, then his treatment of evidential probability would amount to a (broadly) contextualist answer to the second

 © Greco 2016

The basic strategy for saving Bayesian accounts of defeat, if we adopt contextualism about what subjects are entitled to conditionalize on, is to argue as follows. First, admit that in a context in which a subject counts as being entitled to update on *E*, *E* cannot be undermined. So far, this looks like biting the bullet and accepting that certain beliefs are immune to undermining—each context will treat certain beliefs as indefeasible. However, the contextualist can account for a version of the ubiquity of defeasibility in the following way: even if we start in a context in which we count as being entitled to conditionalize on some evidence *E*, and in which *E* thereby counts as being immune to undermining, we can always shift to a new context in which we do *not* count as being so entitled, and in which our support for *E* (if any at all) is treated as defeasible, and may be undermined when we learn some new proposition *E*′.[33] Moreover, it may be that we have a systematic tendency to shift contexts in such a way that claims that in an earlier context are treated as indefeasible are, once the question of their defeasibility is raised and taken seriously, treated as defeasible in a later context.[34]

---

question, since (as mentioned in the previous footnote) he holds that subjects are rationally entitled to update on (and only on) what they know—although as previously mentioned, Williamson (2005) rejects contextualism about knowledge.

[33] Ram Neta (2003, 2004, 2005) has defended contextualism about evidence, and the view I express here is heavily indebted to his work. To my knowledge, however, he has never combined his contextualism with Bayesian approaches to belief updating, or used it to address the apparent challenge to such puzzles presented by the ubiquity of defeasibility. The strategy I pursue in this section also bears some similarity to Lewis's (1996) contextualist strategy for saving closure principles for knowledge from the threat of Kripke's dogmatism paradox. Like Neta, though, Lewis does not apply his discussion of contextualism to Bayesianism.

[34] This may provide a wedge for various species of anti-foundationalism to sneak back in. Suppose our tendency when *E* is challenged is to retreat to a context in which *E* is not foundational but *E*′ is. And suppose our tendency when *E*′ is challenged is to retreat to a context in which *E*′ is not foundational but *E*′ is. Suppose that in the series of claims to which we are disposed to retreat, there are no repeats—i.e. no element in the series *E*, *E*′, *E*′ … is identical to any other. In that case, there is a sense in which infinitism—of the sort defended by Klein (1998, 1999)—would have turned out to be right after all, not as an account of the structure of justification *within* a context, but instead as an account of how we are disposed to shift *between* contexts. Alternatively, if there *were* repeats in the series, then a certain sort of coherentism might be vindicated. I will not pursue these questions here. I am more interested in exploring the consequences of the view under discussion than in deciding whether it should be thought of as a version of foundationalism or not.

 © Greco 2016

The strategy I am suggesting for the Bayesian will need to have two components:

(1) A account of how particular contexts do treat certain beliefs as indefeasible;

(2) an explanation of why no particular belief we consider ever seems to be indefeasible (and so, derivatively, an explanation of our temptation to endorse the claim that all beliefs are defeasible).

While developing these components in depth is beyond the scope of this paper, I hope the following will give the reader some idea of how the story might go, as applied to the cases of defeat discussed earlier in this section.

As for the first component of our strategy, the short answer is that in any context, certain claims will count as legitimate to update on, and, for reasons we have already seen, what you update on will be immune to undermining in the Bayesian framework. For instance, we can agree in the initial case that you are entitled to conditionalize on *Intention*, and that doing so provides support for *Pass*, but that this support is defeated when you learn *Ad Campaign*. The initial, flat-footed Bayesian account of what's going on when you read one story about what the voters intend and then another story about how those intentions are likely to change, can remain untouched.

What about the next case, where the second story concerns flawed polling methodology? If we started out discussing the earlier case, then bringing up this more complicated one is liable to induce a shift into a new context, in which you no longer count as having been entitled to conditionalize on *Intention* in the first place. In this new context—one in which we are taking seriously possibilities in which the polls are unreliable—all you are entitled to conditionalize on is *Polls*, the claim that the polls reported that the majority of Salemites intend to vote 'Yes'. While this claim on its own supports *Pass*, together with *Flawed Source*—which we still take you to be entitled to conditionalize on—it does not.

The next step is similar. When we move to the case in which the second story reports that the first story misreported the poll results, we induce yet *another* context shift. Now that we do not regard you as having unproblematic access to what the polls said, we treat you as not entitled to conditionalize on claims about what the polls reported. Rather, we will think you are only entitled to conditionalize on

 © Greco 2016

claims about what the newspaper *said* the polls said, and this claim will not support the hypothesis that the proposition will pass when combined with the evidence you get from the second story.

Whenever we regard a claim as having been defeated, we will give the same formal account: we will regard somebody as having had an initial body of evidence E that supported some hypothesis H, but as having a final body of evidence E'—a superset of E—that no longer supports E. That is, defeat will always be explained as involving only evidence *accretion*, and never evidence *deletion*. The role that context plays, however, is in determining what gets to count as the initial evidence; rather than giving a context-independent, univocal answer to the question of what a subject's evidence is, we allow that our answer to questions about what a subject's evidence is will change as our context changes (for example, we may initially regard you as entitled to update on claims about what the majority of Salemites intend, but we also recognize that if our context shifts, we may no longer regard you as so entitled).

Just what induces these context shifts? This question is closely related to the demand for the second component of our strategy, the demand for an explanation of why no particular belief we consider ever seems to be indefeasible. The answer will depend on the particular form our contextualism takes, and things will of course look different if we instead opt for some version of relativism or expressivism. But certain basic factors may be common to each approach in the family—perhaps subjects will generally not count as entitled to update on some proposition when there are salient possibilities[35] in which they are in error as to the truth of that proposition. Most of the time, we are ignoring lots of error possibilities, and treat subjects as entitled to conditionalize on claims about their external environment. But in special cases where certain error possibilities are raised to salience, we retreat, and treat subjects as entitled to conditionalize only on other claims—perhaps claims about their experiences, or claims about their inclinations to believe claims about their experiences, or something else. And there is no principled stopping point to this process of retreat—there is no sort of evidence statement about which no doubts could be raised, though in practice we're usually not inclined to retreat very far.[36]

[35] Salient to whom? That may differ depending on the approach

[36] Moreover, the 'retreat' need not always be to weaker and weaker propositions. Rather, as the Neurath's raft metaphor suggests, it might simply be to *different* propositions.

 © Greco 2016

We can use these observations to provide the basis for a general story about why no beliefs seem to be indefeasible: raising the question as to whether some belief is defeasible essentially amounts to an invitation to consider possibilities in which the belief might be false, so once the indefeasibility of a belief is in question, the context will typically shift so as to ensure that the belief is defeasible. That is, once we are considering possibilities in which we are in error as to whether *E*, it will seem irresponsible to simply take *E* for granted; rather, we'll be inclined to accept *E* only if it receives adequate support from other claims we *do* take for granted. But this is just what it is for us to treat *E* as enjoying merely defeasible support. Indefeasibility is 'elusive', in the sense of Lewis (1996); raising the question of whether some claim *E* enjoys indefeasible support will typically bring us into a new context in which *E* enjoys merely defeasible support, even if it initially enjoyed indefeasible support in our old context.[37]

The apparent difficulty for Bayesian accounts of defeat required assuming that the Bayesian must give a single, context-independent answer to the question of what you should conditionalize on when you read a particular newspaper story. As we saw before, whatever answer she gives, there will be cases she cannot handle. But now we see that the Bayesian can reject the demand to tell us, once and for all, what you should conditionalize on in such a case. In most contexts, where we're not worried about certain error possibilities, there is no obstacle to allowing that you are entitled to conditionalize on the content of what the article reports (rather than just, for example, the claim *that* it reported what it did). But other contexts will demand other treatments, and we need not allow that there is some privileged, stable context that we could never be pushed out of—we need not allow that there is any answer to the question of what you should conditionalize on that we could not come to reject after yet another context shift.[38] The resulting position can be summed up, more or less, by the

[37] I should stress that I don't assume that we *always* retreat once a challenge is raised. Some challenges are not taken seriously; attempts to change the context can be resisted. See, for example, Blome-Tillmann (2009).

[38] Earlier in this section I suggested that it would be a defect of a treatment of defeat if it were 'oddly disjunctive'. But one might worry that the sort of account I have suggested on behalf of the Bayesian is just that. We might put the worry as follows: on the story I have been sketching, there are really two quite different types of defeat. The first is the simple sort of defeat that does not involve anything like context shifting. The second is the more complicated sort, which requires a broadly contextualist explanation.

This charge can, however, be resisted. The sense in which the contextualist can deny that she countenances two forms of update (or two associated forms of defeat) parallels the

 © Greco 2016

following remark from Wittgenstein's *On Certainty*: 'The same proposition may get treated at one time as something to test by experience, at another as a rule of testing' (Wittgenstein, 1969, p. 98).

In the absence of more general motivations for contextualist foundationalism, this reply on behalf of the Bayesian might seem ad hoc. But once we admit both that contextualist foundationalism is a viable contender in debates about the structure of epistemic justification, and that Bayesianism is in effect a species of foundationalism in which the beliefs that subjects are entitled to update on play a foundational role, it is hard to avoid the conclusion that we must take seriously contextualist Bayesian foundationalism, according to which the beliefs that subjects are entitled to update on play a foundational role, but which sorts of beliefs get to play that role varies with context. And once we take that view seriously, we see that it has the considerable virtue of providing the foundations for a general account of defeat.

Still, it is worth acknowledging that the sort of contextualism we must appeal to in order to employ the strategy I have been discussing is, in important ways, a departure from some familiar forms of contextualism. Stalnaker (2008, pp. 102–5) distinguishes between *superficial* and *deep* contextualism. Roughly, the sort of context-dependence posited by the superficial contextualist is eliminable, while the sort of context-dependence posited by the deep contextualist is not. The deep contextualist is contextualist 'all the way down', while the superficial contextualist thinks that, once you start digging, you eventually hit some context-independent epistemological bedrock. Without worrying too much about exactly how far 'all the way down' is, we can compare versions of contextualism for 'depth' depending on just how many of our epistemological notions they take to be context-sensitive. For concreteness, it will help to consider some examples.

---

contextualist strategy for affirming 'intellectualism'—the thesis that what one knows depends only on 'truth-conducive factors'. Roughly, while the contextualist will accept different sentences involving epistemological vocabulary (whether those are knowledge attributions or claims about what subjects are entitled to update on) as factors other than truth-conducive factors change (perhaps as the practical stakes change, or as she starts considering new error possibilities), in no context will she affirm that some subject would (or would not) know or justifiably believe some proposition if only the non-truth-conducive factors were different (for instance, if only the practical stakes were different, or if only we were considering fewer error possibilities). See Stanley (2005, pp. 2–3) and DeRose (2009, pp. 24–5), both of whom accept that the sensitive invariantist must deny intellectualism, while the contextualist may accept it. Moreover, I take it that this feature of contextualism is shared by expressivism and relativism, for essentially the same reasons.

 © Greco 2016

Consider the following simple form of contextualism about justified belief. A subject's belief that *P* is properly called 'justified' in a given context only if the epistemic probability of *P* for the subject is above a certain threshold. Which threshold? That depends on the context. The epistemic probability of *P*, however, does *not* depend on context; context only comes into the picture in determining how probable *P* must be in order for a subject's belief that *P* to merit the label 'justified'. The view defended by Cohen (1988) is not *too* far from this.[39] This is a relatively superficial version of contextualism in that it takes there to be a central epistemological notion—epistemic probability—for which contextualism does not hold. Moreover, this sort of superficial contextualism clearly cannot avail itself of the strategy I have been suggesting, since the strategy I have been suggesting requires that epistemic probability *itself* is context-sensitive; in contexts where *E* counts as evidence, we may strictly conditionalize on it, so it will have maximal probability. In other contexts, it will not. So the sort of contextualism Bayesianism requires to deal with the ubiquity of defeasibility goes deeper than at least one influential and familiar species of epistemological contextualism.

David Lewis (1996) defends another version of contextualism, which also turns out to be insufficiently deep for the Bayesian's needs. According to Lewis, knowing that *P* requires eliminating all contextually relevant counterpossibilities to *P*. His view is contextualist, because the set of relevant counterpossibilities varies with context. But what does the eliminating—one's evidence—does not.[40] On Lewis's view, one's evidence is always given by the content of one's experience and one's memory (Lewis 1996, p. 425). If we put things in Bayesian terms, this amounts to the claim that one is always entitled to update on the content of one's experience and memory, no more and no less—and Lewis is no contextualist about the contents of experience and memory. For my purposes, this is a crucial juncture at which Lewis's contextualism takes a superficial rather than a deep form, crucial because the Lewisian contextualist cannot appeal to my

[39] Cohen advances a contextualist version of the relevant alternatives theory of knowledge: on his view, epistemic probability is taken as fixed independently of context, and how probable an alternative has to be before it counts as relevant is determined by context. Also, while they are sensitive invariantists rather than contextualists, Fantl and McGrath (2009) employ a similar strategy on which epistemic probability is stakes-invariant, but the cut-offs for knowledge and justification are not.

[40] Stalnaker (2008, pp. 102–5) gives further reasons why Lewis's view is best thought of a superficial form of contextualism.

 © Greco 2016

strategy for reconciling Bayesian accounts of updating with the apparent ubiquity of defeasibility; Lewis's view straightforwardly implies that our knowledge of the contents of our experience and memory is indefeasible in every context, so if we want to reject that assumption we'll need to look elsewhere.

Most writers who have defended versions of contextualism have only officially committed themselves to relatively superficial forms of contextualism, since they have only defended contextualism about one or another epistemological notion. Moreover, in so far as they invoke some context-*in*sensitive epistemological machinery in explaining their position—such as a context-insensitive notion of epistemic probability—they are committed to rejecting deep forms of contextualism.[41] Still, while deeper forms of contextualism are less familiar in epistemology than their more superficial cousins, I think they deserve to be taken seriously.[42] Where relatively superficial forms of contextualism are unable to reconcile Bayesian accounts of defeat with the ubiquity of defeasibility, deeper forms can.

4. Conclusion

The idea that contextualism can provide a sort of middle way in the debate between coherentism on the one hand, and traditional foundationalism on the other, is not new. But with the exception of some discussion of traditional sceptical problems, there has been relatively little work done exploring how *other* epistemological debates look once we take this middle way.[43]

[41] As an aside, I am inclined to think that many versions of superficial contextualism are not particularly interesting qua epistemological theses (though they may be interesting qua semantic theses); to the extent that the contextualist allows a realm of 'pure', context-insensitive epistemology, epistemologists can largely ignore context and focus their investigations on topics to which it is not relevant. Moreover, the existence of a set of context-insensitive epistemological questions limits the scope of the epistemological work that can be done by contextualist theses. For instance, if we try to solve sceptical problems by appeal to some version of superficial contextualism, such problems will likely reappear concerning epistemological notions where we concede that contextualism will not help. If we are contextualists about knowledge but not epistemic probability, then even if our contextualism helps us explain why utterances of 'I know that I have hands' might be true in ordinary contexts, it will not help us explain why the epistemic probability that I have hands is any higher than the epistemic probability that I am a handless brain in a vat.

[42] See Stalnaker (2008) for further motivation.

[43] Much of Michael Williams's work involves using a sort of contextualist foundationalism to argue that global sceptical challenges are somehow illegitimate. He has not, to my knowledge, discussed the relationship between contextualist foundationalism and Bayesianism. Ram

 © Greco 2016

While I have focused on questions about defeat and belief updating, I hope to have also given the flavour of a general strategy we might use to explore further potential applications of contextualist foundationalism. First, take some view that might seem inextricably linked to traditional versions of foundationalism—versions that require a fixed cognitive home. Then explore whether the idea can be defended while appealing only to mobile homes. The idea that rational updating proceeds by strict conditionalization is one example where I have argued this strategy is fruitful. But I suspect there are others. For instance, there has been much recent work on the relationship between binary epistemological notions (belief, knowledge) on the one hand, and graded ones (credence, epistemic probability) on the other.[44] Many of the writers working on this topic presuppose that a certain flat-footed account of the relationship—on which belief requires credence 1, and knowledge requires epistemic probability 1—cannot be right, since they assume that few if any propositions are believed to degree 1, or have maximal epistemic probability. They take the flat-footed view to be incompatible with the post-Cartesian realization that *all* claims are, at least to some degree, doubtful. But the perspective of the present paper suggests a different way of accommodating that idea. While it may be that no propositions have maximal probablity in *every* context, it can still be that many propositions have maximal probability in *some* contexts. Once we're happy thinking of graded notions like credence and probability as themselves context-sensitive, the obstacle to the flat-footed approach to the relationship between binary and graded epistemological notions may very well dissolve.[45]

G. K. Chesterton (1910) famously claimed that Christianity had not been tried and found wanting, but instead had never been truly tried. Similar claims have been made in defence of communism, the war in Afghanistan, the Reagan revolution, and various other ideas that have, in fact, been tried. Nevertheless, this is the plea I offer on behalf of contextualist foundationalism. Because it promises to illuminate debates about Bayesianism, defeat, and perhaps many other topics as

---

Neta's work (2003, 2004, 2005) also involves something like contextualism about which beliefs are foundational, but to my knowledge he has not applied his version of contextualism to the topic of defeat.

[44] For example, Buchak (2014), Ross (2014) and Leitgeb (2014).

[45] See Clarke (2013) and Greco (forthcoming) for arguments along these lines, though it is not quite clear to me whether Clarke's view should be thought of as a species of contextualism, rather than a species of interest-relative or sensitive invariantism, in the sense defended by Hawthorne (2004), Stanley (2005), and Fantl and McGrath (2009).

 © Greco 2016

well, before we conclude that it is an idea whose time has come and gone we should give it a real try.[46]

References

Aikhenvald, Alexandra Y. 2004: *Evidentiality*. Oxford: Oxford University Press.

Alston, William P. 1976: 'Has Foundationalism Been Refuted?' *Philosophical Studies*, 29, pp. 287–305.

Annis, David B. 1978: 'A Contextualist Theory of Epistemic Justication'. *American Philosophical Quarterly*, 15, pp. 213–19.

Blome-Tillmann, Michael 2009: 'Knowledge and Presuppositions'. *Mind*, 118, pp. 241–94.

BonJour, Laurence 1985: *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

——1999: 'Foundationalism and the External World'. *Philosophical Perspectives*, 13, pp. 229–49.

Buchak, Lara 2014: 'Belief, Credence, and Norms'. *Philosophical Studies*, 169, pp. 285–311.

Carnap, Rudolf 1967: *The Logical Structure of the World and Pseudoproblems in Philosophy*. Translated by Rolf A. George. London: Routledge and Kegan Paul.

Chesterton, G. K. 1910: *What's Wrong With the World*. New York: Cosimo Classics, 2007.

Chrisman, Matthew 2007: 'From Epistemic Contextualism to Epistemic Expressivism'. *Philosophical Studies*, 135, pp. 225–54.

Christensen, David 1992: 'Confirmational Holism and Bayesian Epistemology'. *Philosophy of Science*, 59, pp. 540–57.

Clarke, Roger 2013: 'Belief Is Credence One (In Context)'. *Philosophers' Imprint*, 13, pp. 1–18.

Cohen, Stewart 1987: 'Knowledge, Context, and Social Standards'. *Synthese*, 73, pp. 3–26.

——1988: 'How to be a Fallibilist'. *Philosophical Perspectives*, 2, pp. 91–123.

——1995: 'Is There an Issue About Justified Belief?' *Philosophical Topics*, 23, pp. 113–27.

[46] For helpful comments and discussion, thanks to David James Barnett, Jessica Brown, Sinan Dogramaci, Brian Hedden, Sophie Horowitz, Jack Marley-Payne, Ram Neta, John Pittard, Agustín Rayo, Damien Rochford, Miriam Schoenfield, Robert Stalnaker, Jason Stanley, Jonathan Vogel, Roger White, the Corridor reading group, audiences at Edinburgh and Rutgers, and two anonymous referees at *Mind*.

 © Greco 2016

DeRose, Keith 1995: 'Solving the Skeptical Problem'. *Philosophical Review*, 104, pp. 1–52.
——2009: *The Case for Contextualism: Knowledge, Skepticism, and Context*, Vol. 1. Oxford: Oxford University Press.
Fantl, Jeremy, and Matthew McGrath 2009: *Knowledge in an Uncertain World*. Oxford: Oxford University Press.
Field, Hartry 2009: 'Epistemology Without Metaphysics'. *Philosophical Studies*, 143, pp. 249–90.
Garber, Daniel 1983: *'Old Evidence and Logical Omniscience in Bayesian Confirmation Theory'*. In John Earman (ed.), pp. 99–131. Minneapolis, MN: University of Minnesota Press.
Goodman, Nelson 1983: *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.
Greco, Daniel forthcoming: 'How I Learned to Stop Worrying and Love Probability 1'. Forthcoming in *Philosophical Perspectives*.
Harman, Gilbert 1986: *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.
——2001: 'General Foundations versus Rational Insight'. *Philosophy and Phenomenological Research*, 63, pp. 657–63.
Hawthorne, John 2004: *Knowledge and Lotteries*. Oxford: Oxford University Press.
Ichikawa, Jonathan Jenkins 2013: 'Basic Knowledge and Contextualist "E = K"'. *Thought: A Journal of Philosophy*, 2, pp. 282–92.
Kelly, Thomas, and Sarah McGrath 2010: 'Is Reflective Equilibrium Enough?' *Philosophical Perspectives*, 24, pp. 325–59.
Klein, Peter 1998: 'Foundationalism and the Infinite Regress of Reasons'. *Philosophy and Phenomenological Research*, 58(4), pp. 919–25.
——1999: 'Human Knowledge and the Infinite Regress of Reasons'. *Noûs*, 33, pp. 297–325.
Kotzen, MatthewMS: 'A Formal Account of Evidential Defeat'. Unpublished manuscript.
Leitgeb, Hannes 2014: 'The Stability Theory of Belief'. *Philosophical Review*, 123, pp. 131–71.
Lewis, David 1996: 'Elusive Knowledge'. *Australasian Journal of Philosophy*, 74, pp. 549–67. Reprinted in Lewis 1999.
——1999: *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.
MacFarlane, John 2011: 'Epistemic Modals Are Assessment-Sensitive'. In Andy Egan and Brian Weatherson (eds.), *Epistemic Modality*. Oxford: Oxford University Press.

 © Greco 2016

——2011: 'Relativism and Knowledge Attributions'. In Sven Bernecker and Duncan Pritchard (eds.), *The Routledge Companion to Epistemology*. London: Routledge.

McGrew, Lydia 2013: 'Jeffrey Conditioning, Rigidity, and the Defeasible Red Jelly Bean.' *Philosophical Studies*, 168, pp. 569–82.

Mayo, Deborah 1996: *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Neta, Ram 2003: 'Contextualism and the Problem of the External World'. *Philosophy and Phenomenological Research*, 66, pp. 1–31.

——2004: 'Perceptual Evidence and the New Dogmatism'. *Philosophical Studies*, 119, pp. 199–214.

——2005: 'A Contextualist Solution to the Problem of Easy Knowledge'. *Grazer Philosophische Studien*, 69, pp. 183–206.

Plunkett, David, and Tim Sundell 2013: 'Disagreement and the Semantics of Normative and Evaluative Terms'. *Philosophers' Imprint*, 13(23), pp. 1–37.

Pryor, James 2013: 'Problems for Credulism'. In Chris Tucker (ed.), *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism*. Oxford: Oxford University Press.

Ross, Jacob, and Mark Schroeder 2014: 'Belief, Credence, and Pragmatic Encroachment 1'. *Philosophy and Phenomenological Research*, 88, pp. 259–88.

Schaffer, Jonathan 2004: 'From Contextualism to Contrastivism'. *Philosophical Studies*, 119, pp. 73–104.

——2007: 'Knowing the Answer'. *Philosophy and Phenomenological Research*, 75, pp. 383–403.

Schwitzgebel, Eric 2011: *Perplexities of Consciousness*. Cambridge, MA: MIT Press.

Smithies, Declan 2012: 'Moore's Paradox and the Accessibility of Justification'. *Philosophy and Phenomenological Research*, 85, pp. 273–300.

——2012: 'A Simple Theory of Introspection'. In Declan Smithies and Daniel Stoljar (eds.), *Introspection and Consciousness*. Oxford: Oxford University Press.

——2013: 'On the Unreliability of Introspection'. *Philosophical Studies*, 165, pp. 1177–86.

Stalnaker, Robert C. 2008: *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

Stanley, Jason 2005: *Knowledge and Practical Interests*. Oxford: Oxford University Press.

 © Greco 2016

Stine, Gail 1976: 'Skepticism, Relevant Alternatives, and Deductive Closure'. *Philosophical Studies*, 29, pp. 249–61.
Vogel, Jonathan 1990: 'Cartesian Skepticism and Inference to the Best Explanation'. *Journal of Philosophy*, 87, pp. 658–66.
Wagner, Carl 2013: 'Is Conditioning Really Incompatible With Holism?' *Journal of Philosophical Logic*, 42, pp. 409–14.
Weisberg, Jonathan 2009: 'Commutativity or Holism? A Dilemma for Conditionalizers'. *British Journal for the Philosophy of Science*, 60, pp. 793–812.
——2015: 'Updating, Undermining, and Independence'. *British Journal for the Philosophy of Science*, 66, pp. 121–59.
Williams, Michael 1977: *Groundless Belief: An Essay on the Possibility of Epistemology*. Second edn. 1999. Princeton, NJ: Princeton University Press.
——2013: 'Skepticism, Evidence and Entitlement'. *Philosophy and Phenomenological Research*, 87, pp. 36–71.
Williamson, Jon 2010: *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.
Williamson, Timothy 2000: *Knowledge and Its Limits*. Oxford: Oxford University Press.
——2005: 'Contextualism, Subject-Sensitive Invariantism and Knowledge of Knowledge'. *Philosophical Quarterly*, 55, pp. 213–35.
Wittgenstein, Ludwig 1969: *On Certainty*. Oxford: Basil Blackwell.

© Greco 2016

Paternalism and our Rational Powers

Michael Cholbi
California State Polytechnic University, Pomona
mjcholbi@cpp.edu

According to rational will views of paternalism, the wrongmaking feature of paternalism is that paternalists disregard or fail to respect the rational will of the paternalized, in effect substituting their own presumably superior judgments about what ends the paternalized ought to pursue or how they ought to pursue them. Here I defend a version of the rational will view appealing to three rational powers that constitute rational agency, which I call recognition, discrimination, and satisfaction. By appealing to these powers, my version of the rational will view can rank the wrongfulness of paternalistic acts in terms of (a) the extent to which such acts amount to supplanting the paternalized individual's identity as a rational agent with that of the paternalist, and (b) the degree of mistrust in the paternalized individual's rational agency shown by the paternalistic act. My rational powers account thus provides a more complete account of why paternalism is a powerful, but not decisive or absolute, objection to an act or policy. My rational powers account also provides powerful explanations of why rational suasion deflects charges of paternalism; why consenting to intercessions in one's rational agency negates paternalism; why strong paternalism is generally more objectionable than weak paternalism; and why hard paternalism is generally more objectionable than soft paternalism.

1. Paternalism and our Rational Powers

Paternalism involves interfering with another person, against that person's will, in order to benefit that person in some way. This standard characterization nevertheless leaves a philosophically crucial question unaddressed: With what capacities or aspects of a person does a paternalistic act or policy interfere? Mill's famous description of his harm principle (Mill 1859, part I, ¶9) offers two competing answers to that question.

On the one hand, Mill claims that others are not 'warranted, individually or collectively' in 'interfering with' a person's 'liberty of action' unless they do so for purposes of 'self-protection', that is, in order to prevent that person from harming others. Interfering with a person's ability to act for her own good is not morally justified. Taken as an expression of opposition to paternalism, Mill's remarks suggest that paternalism targets a person's liberty, understood as the capacity

Mind, Vol. 126 . 501 . January 2017
© Cholbi 2016
doi:10.1093/mind/fzv205 Advance Access publication 30 September 2016

to act as one desires. Let us call this the *liberty view* of paternalism's target.

On the other hand, some of Mill's remarks emphasize the wrongfulness of certain means by which we might move others to act in desirable ways. Compulsion, control, and coercion are ruled out, whereas 'remonstrating', 'reasoning', 'persuading', and 'entreating' are permissible means to move others to act. These remarks suggest that the heart of paternalism is not that it interferes with another's liberty as such but that it attempts to influence her behaviour through non-rational causes—that it involves non-rational, or at least less than fully rational, means of influence. The paternalist's target is thus not individual liberty *per se*, but an individual's *rational will* or volition.

Here I will largely set aside the liberty view and concentrate on defending the rational will view of paternalism, according to which paternalism involves wrongful intercession in the operations of a person's rational will. I am not the first to do so; my defence will build on the work of theorists such as Seana Shiffrin (2000), Jonathan Quong (2011, pp. 73-107), and Daniel Groll (2012). Central to their conceptions of paternalism is that paternalists disregard or fail to respect the rational will of the paternalized, in effect substituting their own (presumably superior) judgments about what ends the paternalized ought to pursue or how they ought to pursue them. Paternalistic actors thus attempt to 'take over or control' the rational agency of another (Shiffrin 2000, p. 216). Proponents of the rational will view readily agree with the liberty view that paternalism often interferes with individuals' liberty. Indeed, legal paternalism in particular tends to have this characteristic (Quong 2011, pp. 75-6). However, on the rational will view, interferences with a person's liberty are paternalistic only when (and because) they are also unauthorized intercessions in the operations of their targets' rational will.

My defence of the rational will view proceeds as follows. §1 describes the core commitments of the rational will view and shows how the rational will view elegantly accounts for the role of rational suasion and consent *vis-à-vis* paternalism, as well as accounting for certain otherwise puzzling categories of paternalistic action.

§2 provides the foundation for addressing an issue that adherents of the rational will view have by and large not engaged. That an act or policy is paternalistic should give us moral pause. Yet it would be remarkable if even in cases where enormous benefit could be brought about via paternalistic measures, such measures were impermissible simply by virtue of being paternalistic. Paternalism looks like a

 © Cholbi 2016

powerful, but not necessarily decisive, reason to reject an act or policy. This position—that paternalism is presumptively, but not necessarily all-things-considered, morally impermissible—has a great deal of pre-theoretical appeal.

The rational will view can allow that paternalism has the good or rightmaking feature of advancing its targets' ends or the ends those targets ought to pursue. But the rational will view faces an unappreciated difficulty in explaining how paternalism can be presumptively, but not necessarily all-things-considered, morally impermissible. Paternalistic acts fall into two conflicting moral kinds: They are objectionable because they are intercessions in a person's rational will, but praiseworthy insofar as they are acts of beneficence. Yet a paternalistic act that instantiates the latter kind to a high degree remains objectionable insofar as it belongs to the former kind. The possibility that a paternalistic act might benefit its target to a large enough degree to render it all-things-considered morally permissible seems to require a way of weighing, however inexactly, these two considerations. It might be assumed that such weighing involves only one variable, namely, how much a target is benefitted by a paternalistic intercession. It might be assumed that, where X and Y are two paternalistic acts, X is all-things-considered morally justified if the quantum of benefit *b* to X's target meets or exceeds some threshold, whereas Y is not all-things-considered morally justified if the quantum of benefit *c* to Y's target does not meet that threshold. One of my main aims here is to suggest that this picture of the casuistry of paternalism is too simple, for we cannot assume that X and Y should be assigned the same weight by virtue of being paternalistic. Paternalistic intercessions can engage with our rational wills in different ways. On an intuitive level, a person who tries to encourage my frugality by insisting that I not offer gifts on her birthday has engaged with my rational will in less far reaching and less morally objectionable ways than the person who tries to encourage my frugality by disabling the pop-up advertisements on my computer that contribute to my compulsive spending. A more credible casuistry of paternalism thus requires incorporating and weighing two variables, the magnitude of benefit to the target of a paternalistic act and the specific ways in which that act engages with the target's rational will. The rational will view would thus benefit from an elaboration of the different ways in which paternalistic acts engage with their targets' rational wills.

I provide such an elaboration by spelling out the internal architecture of our rational agency. More specifically, our rational agency is

 © Cholbi 2016

constituted by three rational powers, each of which plays a distinct role in agents' choice and action and each of which can be engaged by the paternalistic actions of others. I argue in §3 that paternalistic intercessions are wrong because and to the extent that they intercede in these rational powers: These powers vary in their centrality to an individual's practical identity, such that the more central a power is to the constitution of an individual's practical identity, the more wrongful are paternalistic interventions that engage that power. §4 offers a second way of conceptualizing paternalism's wrongmaking features in terms of intercessions in our rational powers, namely, that paternalistic actors show varying levels of mistrust in the rational competency of others based on which of these powers such actors intercede in. By grasping the internal architecture of the rational will, we can thus see how the rational will view might normatively weigh the wrongfulness of paternalistic acts against their goodmaking feature (i.e., that, when successful, they aid their targets in securing their ends or ends they ought to seek). My *rational powers account* is thus a version of the rational will view that improves upon it by demystifying the variety of ways in which acts or policies can intercede in the rational wills of others.

I turn to an objection in §5: If we have evidence that a potential target of our paternalistic intercessions is in fact untrustworthy in the exercise of her rational powers, why is it even *pro tanto* wrong to intercede in those powers on her behalf? I defend the notion that respect for rational agency requires deference toward agents even when they do not exercise that agency with full competence or capability.

§6 highlights how my rational powers account explains common philosophical stances about different varieties of paternalism.

2. The rational will view and its explanatory advantages

Rational will views of paternalism share two features. First, rational will views require that paternalistic acts involve the substitution of the paternalist's judgment or agency for that of her target. Every account of the nature of paternalism agrees that paternalists act, or try to induce their targets to act, in ways that reflect what the paternalist contends is good (in however broad a sense) for the target to do. As the paternalist sees it, her target is disposed to choose or act either in self-defeating ways, by failing to act in accordance with what the target

 © Cholbi 2016

antecedently believes to be good or desirable, or in foolish ways, by failing to pursue ends or goods that, if the target were more reflective or knowledgeable, she would pursue. What distinguishes the rational will view is its emphasis on how paternalistic intercessions seek to sever the usual causal relationship between the target's judgment or agency and the target's choices or actions by interpolating the paternalist's own judgment. For adherents of this view, this feature is central to the moral objectionability of paternalism. Seeking to guide another's actions or choices by one's own lights rather than theirs represents an 'intrusion into and insult to [that] person's range of agency' (Shiffrin 2000, p. 218).[1]

Second, rational will views think of paternalism as motivated by the belief that the target of paternalism is, when compared to the paternalistic actor, deficient in some facet of her rational competence or judgment.[2] This second feature is a direct consequence of the first: In order for paternalists to believe that the substitution of their judgment for that of their targets is warranted, they cannot believe, on pain of contradiction, that their own judgment is not superior to their targets' with respect to the matter at hand. Of course, this belief in one's superior competence or judgment need not be global: A may believe herself superior to B with respect to financial decisions but not with respect to dietary ones, and so be inclined to financial paternalism but not dietary paternalism, for instance. Moreover, this need entail neither that the paternalistic actor's judgment is impeccable nor that the target's judgment is incompetent. A may believe that, while she is merely competent with respect to financial decision making, B is a complete financial neophyte who would benefit from her paternalism

[1] As I see it, the intrusiveness of paternalism is normatively fundamental and the insultingness derivative, that is, paternalism insults others because it intrudes upon others' rational agency, not vice versa. Cornell (2015) appears to disagree, offering an account of paternalism in which its insulting 'expressive content' is fundamental. Others who emphasize that paternalism is insulting or demeaning include Kleinig (1983), Feinberg (1986, pp. 23-4), and Anderson (1999).

[2] Groll (2012, p. 718) presents his version of the rational will view in terms of the target being 'likely to fail to exercise a capacity for sound judgement in the situation at hand'. Quong (2011, pp. 80, 83) argues that paternalism is better described in terms of a 'negative judgement' about the target's abilities than as a judgement that the target's abilities are inferior. His chief reason for this is that paternalistic actions sometimes depend on the belief that their targets lack willpower, rather than poor judgement, and in these cases, the paternalizer need not think that her willpower is superior to her targets. I do not share Quong's assumption that lack of willpower is not an instance of judgement, and while I agree that the paternalizer need not believe his willpower is superior, his judgement is putatively superior in the sense that he believes himself better situated to enable his target to realize her chosen ends.

 © Cholbi 2016

(whereas C, a financial expert, would be in a position to act paternalistically with respect to B and to A). Furthermore, the rational will view does not entail that the paternalist literally sees her target as childlike or infantile. She merely takes her target to be less competent in her rational judgment or competency than she.[3]

So presented, the rational will view understands the wrongfulness of paternalism as a complex wrong with both comparative and substantive aspects (Quong 2011, pp. 100-2). First, paternalists treat their targets as if they are, with respect to the choices in which the paternalist intercedes, inferior to them. Second, their treating their targets in this way is manifest in intercessions in the operations of their agency or volition. As I see it, neither the comparative nor the substantive aspects of paternalism are more morally basic, and they cannot be disentangled without losing sight of the moral essence of paternalism. Intercessions into other agents' rational volition need not be paternalistic. Lying, arguably, intrudes into others' rational volition but is often not paternalistic. Similarly, there are other ways to treat others as inferior without being paternalistic. But in concert, these two aspects underscore how the moral essence of paternalism is to deny its targets their status as equal rational participants in social life, thereby impugning, as Rawls (2005, p. 19) puts it, their 'capacity to form, to revise, and rationally to pursue a conception of [their] rational advantage or good'.

Furthermore, the rational will view's conception of paternalism's wrongfulness need not exhaust the ways in which paternalistic acts can be wrong. Indeed, one complication in evaluating the moral seriousness of paternalistic acts is that the vast majority of such acts also fall into other morally significant kinds. Most obviously, paternalistic acts are beneficent insofar as they aim at their targets' good. Conversely, many paternalistic acts are also wrongs of another kind. They are acts of violence, promise breaking, coercion, deception, manipulation, intrusions into privacy, etc.[4] As a result, it can prove difficult to distil the distinct contributions that paternalism makes to the all-things-considered moral properties of a paternalistic act. For instance, to what degree is an act of paternalistic coercion wrong because it is coercive and to what degree is it wrong because it is paternalistic? I make no pretence of answering such questions with any precision here. My

[3] De Marneffe (2006), for one, exaggerates Shiffrin's account in this regard.

[4] I hesitate to say that *all* paternalistic acts also fall into other act kinds that are typically wrong. Paternalistic offers, for example, seem to lack any other morally troubling properties.

 © Cholbi 2016

main objective is to defend the rational will view as the correct account of the bad or wrongmaking feature of paternalism, and it would be far too ambitious to offer a comprehensive casuistry of paternalistic acts. However, I will later appeal to readers' intuitive responses about the moral wrongfulness or objectionability of different paternalistic acts, and there is the very real danger that these responses will be coloured by the fact that these acts also belong to other morally objectionable kinds. I make these remarks to urge readers to focus on the paternalistic features of these acts and to set aside, to the extent possible, other good or badmaking features of these acts.

The rational will view appears to understand paternalism as a psychological kind, as a set of attitudes that paternalists bear toward their subjects. However, this psychological nature of paternalism can be easily misunderstood. That paternalists necessarily have a certain set of attitudes toward their subjects implies neither that paternalists know they have these attitudes nor that they act under descriptions referring to those attitudes. Indeed, in an important sense, paternalism is always well intentioned, inasmuch as it seeks to make its subjects better off. But this is wholly compatible with paternalists failing to recognize how their paternalistic acts rest on judgments of their subjects' relative inferiority or how those acts intercede in their subjects' rational volition. Hence, paternalism is perhaps best understood as an action-theoretical kind, akin to 'cruelty'. Cruelty is most readily characterized by referring to the psychology underlying cruel acts, that is, as the wilful and knowing subjection of another to pain or distress. But characterizing cruelty in this way does not preclude there being unintentional cruelty, accidental cruelty, cruelty done through ignorance, etc. Similarly, I have characterized paternalism psychologically, in terms of the judgments and attitudes of paternalistic actors, even though paternalism, like cruelty, can be accidental, unintentional, done through ignorance, and the like. In this case, it proves perspicuous to characterize the nominative (paternalism, or paternalistic acts) in terms of the adverbial (acting paternalistically).[5] Moreover, if paternalism is an action-theoretical kind, then it is possible to ask relevant moral questions about 'impure' instances of paternalism. Just as we may ask whether, say, unintentional cruelty is as blameworthy as intentional cruelty, we can ask whether unintentional paternalism is as

[5] I take my claim that paternalism is an action-theoretical kind to be in the spirit of the proposal in Grill (2007) that paternalism rests on the 'invocation of certain reasons for certain actions'.

 © Cholbi 2016

blameworthy as intentional paternalism. I do not address such questions here. I merely point out that understanding paternalism as an action-theoretical kind enables us to differentiate between, and morally interrogate, instances of paternalism whose psychological underpinnings differ from the paradigm instances of paternalism.

Understanding paternalism as an action-theoretical kind also enables the rational will view to account for two sorts of paternalistic acts that might otherwise seem puzzling.

First, the rational will view can allow for paternalistic omissions (Dworkin 2013a, p. 29ff.). The simplest example is a paternalist withholding information from her target for the latter's benefit. Admittedly, such withholding does not interfere with the target's rational deliberation inasmuch as the target will deliberate, and presumably choose and act, just as she would have. This act does not disrupt its target's extant trajectory of rational deliberation. However, all paternalistic reasoning operates with counterfactuals: The paternalist reasons that by interceding in her target's rational agency, the target would (or will) choose or act differently—and better. According to the rational will view, these counterfactuals will sometimes concern what one's target will do if one *omits* some action, thereby ensuring that the target will continue to deliberate, choose, and act as the paternalist expects. But the rational will view recognizes no basis for normatively distinguishing between these counterfactuals and counterfactuals about what one's target will do if one *performs* some action, thereby attempting to guide the target's deliberation, choice, and action in the ways the paternalist believes are better for the target. In either case, the paternalist seeks to regulate her target's rational agency because she believes her own agency superior.

In a similar vein, the rational will view can also explain how offers can be paternalistic. For example, suppose A, believing B is inferior to her with respect to his rational capacities, bribes B to do X because A believes B ought to do X for B's own benefit. This is an instance of paternalism because the reasons A's bribe provides B for doing X do not relate to the reasons B may have had for X prior to the bribe. A has given B *a* reason to act as A believes B ought to act, but assuming that the choice B must make is a typical one, this is the wrong kind of reason, a state-given rather than an object-given reason, to use Derek Parfit's nomenclature (2001). A has in effect made it the case that B ought to do X but not for the reasons that merit B's doing X. Admittedly, A has not circumvented B's rational volition. Yet a condition of A's showing respect for B's rational capacities is that A

 © Cholbi 2016

rationally persuades B to do X on the basis of X's independent merits. Such paternalistic offers may be all-things-considered justified, but like some instances of rational persuasion, they nevertheless exhibit the disregard for another's rational capacities that morally speaks against paternalism.

Let us conclude this section by noting how the rational will view accounts for two widely acknowledged normative features of paternalism.

First, the rational will view provides a simple explanation of why attempting to influence another's choice or conduct through transparently rational means typically deflects the charge of paternalism (Shffrin 2000, p. 213). For even when one's behaviour has other features commonly associated with paternalism (for example, that one aims to induce another person to act in ways that one perceives to be in their interest), to provide reasons in the hope that doing so will move another to act in a particular way indicates a desire to engage his or her rational will. Providing reasons in the sincere hope of persuading another to act as you believe she ought expresses minimal confidence in the other's rational powers and amounts to giving the individual the opportunity to demonstrate her rational competency or judgment. But if a willingness to see if a putative target of paternalism can be moved by sound reasons, and is therefore trustworthy with respect to her rational competency or judgment, is evidence of non-paternalistic intent,[6] then overtaking another's rational will by substituting one's own competency or judgment for the other's turns out to be essential to paternalism.

We should not, however, tie paternalism too closely to the rejection of rational suasion. As I will detail further in §5, paternalism fails to respect the other person *qua* rational agent. But as George Tsai (2014) has recently argued, the mere fact that one provides reasons to another in an effort to persuade her to act as one believes she ought does not entail that one shows full respect for her rational agency. In providing others with reasons, we do not ignore their rational capacities nor do we directly substitute our judgments for theirs. But we can provide others with reasons in ways that nevertheless fail to honour their capacities for rational volition. We can, through the timing, tone, etc., of our interactions with others provide them with reasons

[6] I say 'evidence' because the willingness to see if rational suasion will move another to act as one believes she ought is nevertheless compatible with lacking respect for her rational agency. One might simply find rational suasion more convenient, etc., than non-rational, paternalistic means of influencing her choices and conduct and so attach no special importance to her rational agency.

 © Cholbi 2016

whilst nevertheless mistrusting their capacities as rational agents, thus precluding their exercising their decision making capacities in ways that are robust, authentic, and independent. For example, inundating another agent with reasons relevant to a particular choice while subtly pressuring him to make a speedy choice can make it difficult for that agent to deliberate solely by reference to the reasons germane to that choice. Tsai shows that there is more to respecting others' rational capacities—and hence more that is necessary to avoid paternalism—than the provision of reasons. We must allow them to engage fully with those reasons on their own terms, to work out for themselves how such reasons bear on their choices. Only then can others exercise their rational agency competently and authentically.[7] In order that we not be paternalistic in trying to persuade another, we must rationally persuade her (as Aristotle might have said) in the right manner, at the right moment, and in the right spirit. Indeed, when we act paternalistically, it can be said that our acts fail to exhibit the *virtue* of respect.

Secondly, the rational will view offers an elegant explanation as to why another's consent deflects charges of paternalism. When another consents to our treating her in a particular way, we may still harbour the belief that we have better judgment than she. Yet so long as a person's consent to our conduct is minimally rational, our treatment of that person neither fails to engage her will nor amounts to literally substituting our judgment for hers, since her consent provides the judgment that invalidates a charge of paternalism.[8] Consent directly undermines central features of paternalism as the rational will view understands it. The rational will view can thus appeal to a condition or property—that paternalism fails to engage another's rational agency—intrinsic to paternalism itself in order to explain why persuading another or acting with her consent obviates the charge of paternalism.

3. Three rational powers

It might appear that when an instance of paternalism is morally justified, this is only because its target is benefitted to a certain degree.

[7] See Christman (2014) for an elaboration of competency and authenticity as conditions of autonomy.

[8] Such consent may not invalidate other charges levelled at such conduct (for example, that it is manipulative, deceptive, etc.). Furthermore, the presence of consent makes it possible to consent to being subject to force, for example, as psychiatric patients do when then consent to Ulysses contracts regarding their treatment while incompetent, without such force being rejected as paternalistic.

 © Cholbi 2016

That model, I shall suggest here, is too simple to ground an adequate casuistry of paternalism. According to the rational will view, the wrongmaking feature of paternalism is that paternalism amounts to denying equal moral status to another because it intrudes into her rational agency in ways that indicate judgments of inferiority. This wrongmaking feature comes in degrees, I shall argue, because paternalistic acts form a heterogeneous class. Paternalistic acts differ, I shall argue, with respect to which of the rational powers that constitute our rational agency they intercede in. In seeking the rational powers that constitute rational agency, we are seeking the capacities necessary for an individual to choose, intend, and act on reasons whose authoritativeness or reasonableness she recognizes. The rational powers are therefore those capacities needed for practical deliberation and choice to be possible. They are what we possess 'just in virtue of being an agent' (Shiffrin 2000, p. 219).

Here I will focus on those rational powers unique to practical, as opposed to theoretical, rational agency. In his discussion of the value of individuality, Mill asserts that those who do not aspire to individuality, those content to let the world choose their plans of life for them, need nothing more than the power of 'ape-like imitation'. But those seeking to forge their own paths in life must possess certain rational powers:

> He who chooses his plan for himself, employs all his faculties. He must use observation to see, reasoning and judgment to foresee, activity to gather materials for decision, discrimination to decide, and when he has decided, firmness and self-control to hold to his deliberate decision. And these qualities he requires and exercises exactly in proportion as the part of his conduct which he determines according to his own judgment and feelings is a large one. (Mill 1859, ch. 3, ¶4)[9]

Here Mill highlights three of the powers necessary for individuals to be practically rational agents.

First, in order for practical deliberation and choice to commence, we first 'gather materials for decision'. An agent must begin her deliberation from some menu, however small, of options that she perceives as in some respect minimally choiceworthy. We might call these options reasons, considerations, concerns, etc. For ease of exposition, I will call these options *ends*. Not every agent perceives every potential

[9] Quong (2011, p. 81) similarly claims that the relevant abilities about which the paternalist has a negative judgement are her targets' 'practical reasoning, willpower, and emotion management'.

 © Cholbi 2016

end as minimally choiceworthy. This can be due simply to ignorance of it as a possible end. In other cases, an agent can be aware of a prospective end but see nothing intelligibly good about that end, nothing that speaks in favour of it as an object of rational pursuit. Warren Quinn (1993) offers the example of an individual with a desire to turn on every radio she encounters. For most agents, the end served by acting on this desire is, absent further information, simply not intelligible. There is no description of this end under which most agents find its pursuit good in the least. Granted, differences in individual psychology can often explain why one agent finds an end minimally choiceworthy where another does not. And although this capacity to perceive ends as choiceworthy is more a passive power like perception than an active power like choice, it is a rational power in that it involves a sensitivity to those facts that give us reasons to adopt attitudes toward particular objects. Let us call this first rational power, the power to recognize ends as minimally choiceworthy, the power of *recognition*.

Agents with the power of recognition must also be able to choose among these ends and endorse some among them as worthy of being pursued in action. This is a rational power in two senses. First, agents sometimes must weigh competing or incompatible ends in order to engage in rational choice or action. Second, in order for choice to be rational, it must respond to the features of the ends chosen that render those ends choiceworthy. In both of these respects, agents must attend to and attach evaluative significance to various features of ends. Following Mill, let us call this capacity to rationally select, from among those goods one recognizes as minimally desirable or choiceworthy, which goods to endorse and thereby render objects of rational pursuit the power of *discrimination*.

Finally, agents must possess the 'firmness and self-control' necessary to put their choices and intentions into action. Action essentially involves pursuing those means, either causal or constitutive, to realize the ends selected via the power of discrimination. Doing so requires the exercise of instrumental rationality, or as I will call it, the power of *satisfaction*.

Let us fix on a set of examples to illustrate how paternalistic intercessions differ with respect to which of these rational powers are engaged.

C, a recovered alcoholic, is aware of the genetic and environmental factors that make it more likely that children of alcoholics will become problem drinkers. Thus, C desires that her adult child, D, refrain from

 © Cholbi 2016

drinking alcohol. She desires this for D's sake and believes that D's judgment or competency with respect to choosing whether to consume alcohol is weaker than hers. She would very much prefer that D not have to wrestle with alcohol in order to perceive its dangers, as C did. If the above picture of our rational powers is correct, then C has at her disposal several levers that she might pull in order to influence D's rational will in the hope that D will not drink. For reasons that will become apparent shortly, let us confine our attention to paternalistic intercessions in which C lies to or deceives D.

C could try to make it the case that D has no desire at all to drink alcohol. The most decisive way to do so would be for C to eliminate drinking from the menu of ends from which D can choose. Admittedly, it is difficult to imagine how to do this. Perhaps C could undertake a long-term program of 'censorship' directed at D, designed to keep D ignorant of alcohol and its allure. She might dine with D only at restaurants that do not serve alcohol, change the television channel whenever alcohol advertisements are broadcast, interfere with D's personal relationships so that none of D's friends imbibe, etc. If C's censorship campaign succeeds, then drinking will not appear on the menu of minimally choiceworthy ends that D recognizes.

Another way for C to forestall D's desiring to drink alcohol is to 'tilt the scales' against drinking by making it seem relatively less choiceworthy. C could achieve this by wilfully exaggerating the dangers of alcohol to D, thus making D less likely to drink.

Finally, where C is confident that D endorses abstinence as an end, C might take measures to ensure that D prudently pursues that end. C might, for instance, deceive D about whether D's companion, known to be a heavy imbiber, has called to suggest that D join him for an evening at the local watering hole.

The ways in which C might try to prevent D's drinking illustrates how different paternalistic intercessions target different rational powers. When C undertakes the campaign to suppress D's awareness of alcohol and its appeal, C intercedes in D's power of recognition, taking alcohol off the menu of minimally choiceworthy ends that D may choose to pursue. When C exaggerates the dangers of alcohol to D, C intercedes in D's power of discrimination, trying to make it the case that drinking is not an end that D finds sufficiently attractive to pursue, despite D's continuing to recognize alcohol consumption as a minimally choiceworthy end. In deceiving D regarding the invitation to the local watering hole, C intercedes in D's ability to successfully act on his choices and so intercedes in D's power of satisfaction.

 © Cholbi 2016

These three powers are obviously not sufficient on their own for practical agency. As Mill saw, practical agents must also have other perceptual and cognitive capacities in order to exercise their powers of rational choice. But these powers do constitute the distinctly practical or choice-orientated core of powers that a rational agent has.

4. Rational powers and respect for practical identities

With this sketch of the three rational powers in hand, I propose that the rational will view of paternalism, according to which paternalism involves substituting one's judgment for another's and thereby failing to engage another's rational will, ought to conceptualize the wrongs of paternalism in terms of which of these rational powers paternalistic acts seek to supplant. In other words, by understanding how various kinds of paternalism attempt to supplant the three powers necessary for rational volition, we can understand how various paternalistic intercessions are illegitimate based on which of these powers they intercede in. More specifically, this account locates the grounds for opposition to paternalism in truths about individual identity. According to this account, paternalism is a condemnation of our competency as rational agents. But it is also a condemnation of who we are—of important elements of our identity as individuals—and this gives us a special reason for opposing paternalism.

As we noted in §1, one challenge associated with judging the wrongfulness of paternalism is that paternalistic acts are very often also acts of another morally objectionable kind. This fact can make specific judgments about the wrongfulness of paternalistic acts problematic. For argumentative purposes however, the three examples provided in the previous section have the advantage that they are also wrongs of the same kind (or similar kinds), lies or acts of deception. Hence, this set of examples should allow us to set aside the other wrongmaking features paternalistic acts typically have and distil the distinctive paternalistic wrongs involved in each.

Our individual identities are not exhausted by our rational wills. But the rational will dominates the practical identity around which a large portion of interpersonal morality is orientated. When others treat us paternalistically, they judge us inadequate or inferior as rational agents and substitute their judgment for ours on our behalf. Others' paternalism thus colonizes our identities as rational wills. In interceding in the rational powers that constitute our rational agency,

 © Cholbi 2016

paternalists act wrongly because they attempt, perhaps only momentarily, to be us—to operate the levers of our rational agency on our behalf.

But these intercessions colonize our rational wills in different ways, based on which of the rational powers are interceded in.[10] Hence, the precise ways in which paternalists seek to supplant some element of our practical identities varies, and as we might anticipate, the wrongfulness of paternalistic acts is partially determined by how they supplant our practical identities. In fact, the rational powers stand in relations such that intercessions in one power may also effectively supplant others.

As Christine Korsgaard (1996, p. 100) puts it, 'When you deliberate, it is as if there were something over and above all your desires, something which is you, and which chooses which desire to act on'. This standpoint 'over and above' one's desires is one's rational agency or will. The rational will—that something 'over and above' our particular choices and desires—engages with possible objects of choice in order to fashion plans or agendas for action. Competency as a rational agent, and one's identity *qua* rational agent, are thus determined by the relation between one's rational will and the objects of rational choice. Paternalism 'comes between me and my plans for myself' (Conly 2013, pp. 77, 190-2). When C intercedes in D's power of satisfaction by trying to prevent D from joining his companion at the local drinking establishment, C shares D's agenda but doubts that D can successfully pursue that agenda. Hence, C has interceded in D's agenda in the weakest way possible, and though D has reason to object to such an intercession, these reasons are relatively weak because one's power of satisfaction is not a power that is especially definitive of oneself as a distinct practical agent.[11] Given D's agenda to forego alcohol, the choice of how to satisfy this end is essentially the same choice faced by any agent with the same agenda and equipped with the same resources, constraints, etc. By settling all the normatively important facts and settling on his practical agenda, D has already set

[10] I do not intend that an act of paternalism can only intercede in one of these three powers. No doubt some such acts intercede in more than one of these powers.

[11] This claim comes with an important 'all other things being equal' clause: It may be that for some deeply important end of mine (attending the church of my choice), interceding in my power of satisfaction is something I would very strongly object to, perhaps more strongly than I would object to having my power to satisfy some other less important end interfered with (purchasing the brand of pasta of my choice). But I have in mind that different sorts of intercessions in rational volitions are more or less objectionable relative to a given end, not across different ends. All the other claims I make in this section regarding the objectionability of paternalistic acts should be read in a similar, end-relative light.

 © Cholbi 2016

the parameters under which C's intercession can be successful. So though C intercedes in D's agency by interceding in his power of satisfaction, this intercession is relatively incidental to who D is as a rational agent with a particular history of decision and choice. The rational power C supplants while acting on D's behalf, while not trivial in value, is not especially central to D's identity as a distinct rational agent.

However, when C intercedes in D's power of discrimination by lying to D about alcohol's dangers, C seeks to influence D's understanding of the facts that determine D's choice agenda. With respect to this particular agenda of D's, C acts on D's behalf by seeking to shift the weight of reasons D recognizes in favour of not drinking. Notice that even though C intercedes in the power of discrimination, her intercession (if successful) will also shape how D exercises his power of satisfaction. If D ends up believing C's lies about alcohol's dangers and makes refraining from alcohol consumption his agenda, then D must now exercise his power of satisfaction in order to satisfy this new agenda. Whereas effective exercise of D's power of satisfaction previously demanded he ascertain how to continue to drink, now effective exercise of that power demands he ascertain how not to. The power of discrimination thus encompasses the power of satisfaction in that the exercise of the former determines what counts as an effective exercise of the latter. Hence, intercessions in the former are, in comparison with intercessions in the latter, more thoroughgoing colonizations of one's rational powers, and hence amount to more complete takeovers of one's identity as a rational being.

This colonization of one's identity as a rational being is even more extensive when one intercedes in another's power of recognition. Here a successful intercession determines not just the actual agenda a person will enact, but the scope of possible agendas she may enact. Intercessions in the power of recognition are hence intercessions in an agent's 'meta-agenda'. C's attempted censorship campaign bears the same relation to D's power of satisfaction as did C's lying to D about alcohol's dangers. Now D must determine how to refrain from drinking. But in interceding in D's power of recognition, C also shapes what successful exercises of D's power of discrimination look like. Absent C's intercession in D's power of recognition, D's power of discrimination would be directed at drinking as a possible end. With drinking removed from D's deliberative field, D can safely shape his practical agenda without attending to the merits of alcohol. Thus, while the power of discrimination encompasses the power of satisfaction in that

 © Cholbi 2016

the exercise of the former determines what constitute successful exercises of the latter, the power of recognition encompasses both of these, determining the conditions for the successful exercise of both powers.[12]

We thus have reason to morally object to different forms of paternalism depending on which of our rational powers are interceded in, because these powers play different roles in determining how we rationally choose and deliberate and so play different roles in constituting us as rational agents. One might nevertheless believe than any of the three intercessions described in the examples above are morally justifiable despite their being paternalistic. However, if I am correct, the justificatory burden a would-be paternalist faces depends in part on which of the three rational powers she intends to intercede in. The reasons needed to justify paternalistic intercessions in the power of recognition must be stronger than the reasons needed to justify paternalistic intercessions in the power of discrimination, which in turn must be stronger than the reasons needed to justify paternalistic intercessions in the power of satisfaction. Put in contractualist terms, we have stronger reasons to endorse a principle disallowing intercessions in the power of recognition than to endorse a principle disallowing intercessions in the power of discrimination (though of course we have reasons to endorse both such principles). Likewise, we have stronger reasons to endorse a principle disallowing intercessions in the power of discrimination than to endorse a principle disallowing intercessions in the power of satisfaction. Put differently: While our practical identities—the goals, values, or concerns we identify with—are no doubt influenced from many directions, the ultimate determination of our practical identities should largely be left up to us. Paternalism is wrong because, and to the degree that, the paternalist regulates the rational powers through which we fashion our identities.

Note that these claims are issued from the point of view of someone targeted for paternalism, i.e., the reasons we have for rejecting

[12] A referee for this journal suggests that recognition and satisfaction resemble one another in that interceding in these powers involves arranging the world in a certain way, whereas interceding in the power of discrimination is more direct in that the intercession aims to give faulty weights to different possible ends. I grant that interceding in the power of discrimination is more *causally* direct than interceding in the other powers. This observation illustrates that interceding in a person's rational powers can involve the modification of the powers themselves (or their operations) or the modification of the objects ('the world') at which those powers are directed. However, this observation does not undermine the claims I make here about the priority relations among the powers with respect to rational deliberation and agency.

 © Cholbi 2016

paternalistic intercessions in our own choices and actions vary according to which of our rational powers such intercessions engage. For their part, paternalists will presumably intercede in whichever power it is easiest to engage so as to achieve their aims. Hence, even if (for example) a paternalist could benefit her target by interceding in her target's power of discrimination, she may find it easier to achieve that same aim by interceding in the target's power of satisfaction. Indeed, if my claims are correct, a conscientious paternalist will try to minimize her 'colonization' of her target's rational powers by opting to intercede in the least fundamental power compatible with her achieving her beneficent aim.

No matter whether a paternalist engages with the power of recognition, of discrimination, or of satisfaction, the *pro tanto* wrong of paternalism belongs to the same generic kind: She intercedes in those powers so as to lead the agent to choose or act as she believes the agent ought to choose or act, and she does so from the belief that she is more competent than her target with regard to such choices and actions. However, by arraying paternalistic intercessions in terms of which rational powers they engage, the rational powers account provides a richer account of how paternalism is presumptively, but not necessarily, impermissible. Suppose we assume that paternalistic intercessions can be all-things-considered justified if they are likely to augment their targets' welfare (or advance their ends) to some sufficiently high degree. The rational powers account asserts that, for some given quantum of welfare, the fact that a paternalistic act will likely provide that quantum of welfare to its target must be measured against which of the rational powers that act intercedes in. If for example a paternalistic act provides sufficient welfare to its target to justify an intercession in the power of recognition, then it also provides sufficient welfare to justify an intercession in the powers of discrimination or satisfaction (since justifying intercessions in these powers faces less of a burden than justifying intercessions in the power of recognition). The converse, however, will not necessarily hold. An intercession in the power of satisfaction likely to yield a given quantum of welfare to its target may not be great enough to meet the justificatory burdens imposed by the powers of discrimination or recognition. Note that making welfare relevant to the all-things-considered morality of paternalism does not make the rational powers account 'welfarist': Paternalistic acts, just in so far as they are paternalistic, trace their wrongfulness back to which of the rational powers are interceded in, and there is no inherent relationship between these powers and welfare. It is not the case, for

 © Cholbi 2016

example, that intercessions in the power of recognition necessarily contribute more (or detract more) from their targets' welfare than intercessions in the other powers. So too for any other goods that might be realized via paternalism (improvements in moral virtue, say). The goodmaking features that may play a central role in the all-things-considered justifiability of paternalism are therefore logically and normatively independent of paternalism's wrongmaking feature, that it intercedes in others' rational agency or volition.

And again, we must keep in mind that these features in turn logically interact with other morally salient features of paternalistic acts (that they are deceptive, manipulative, violent, etc.) that determine the all-things-considered moral permissibility of such acts. For instance, an act that interferes with a target's power of satisfaction (tying a person to a chair so that she cannot eat a scrumptious dessert) may, from a broader moral perspective, be more objectionable than an act that achieves the same end by interfering with a target's power of discrimination (misleading the target about the dessert's scrumptiousness, say), despite the latter's being more objectionable, *qua* paternalistic act, than the former. How we intercede in another's rational powers can be as morally significant as what powers we intercede in. But this, in my estimation, is what we should expect from an account of paternalism's wrongness: an explanation of why paternalism is a strong reason against an act or policy, but not one that accounts for every moral qualm we might have about paternalistic acts.

5. Paternalism and mistrust

A second way of understanding the wrongmaking feature of paternalism (on my rational powers account) appeals to the varying levels of mistrust that paternalists show when they intercede in a person's rational powers. The rational will view maintains that paternalism's wrongfulness rests on paternalists' manifesting an attitude of superiority with respect to others' rational volition or competency. We desire to be trusted by others, to be seen as rationally competent custodians of our own affairs. Being trusted is a prerequisite for inclusion, equal standing, and the kind of solidarity amidst diversity that characterizes healthy social relations. Conversely, when we fail to trust in one another's rational powers, marginalization, hierarchy, and opacity in our social relations are the likely results.

 © Cholbi 2016

The three rational powers play distinct roles in our practical deliberation, and as a result, those who intercede in these powers show distinct levels of mistrust in our competency as rational agents. Consider first C's intercession in D's power of recognition, suppressing D's awareness of alcohol altogether so that D cannot recognize alcohol consumption as even minimally choiceworthy. Such an intercession profoundly manipulates the terms in which D can subsequently exercise his rational will. By taking alcohol consumption off the menu of minimally choiceworthy ends D might choose, C has sought to preclude D from selecting this as one of his ends. D ends up ignorant of alcohol's charms, and in this case, D's ignorance of the merits of this end may well be bliss. Still, C has concluded that D's judgment in this regard is sufficiently inferior to hers that it is best not even to risk the prospect that D will choose to drink, and as a result C's paternalistic intercession makes it impossible for D to rationally appraise the merits of drinking. D ends up not choosing to drink, but not thanks to having been exposed to the merits of alcohol consumption. Hence, with respect to alcohol consumption, C has substituted her judgment for D's in the most thorough and wide reaching way possible. By removing alcohol consumption from D's deliberative field altogether, C has determined that D's judgment is so impaired that D ought not to deliberate about alcohol's merits at all. C's beliefs thus reflect a *deep* mistrust of the competency of D's rational will or agency.

In contrast, by interceding in D's power of discrimination, C does not remove alcohol consumption from D's deliberative field. Instead, C shapes D's deliberation by modifying D's reasons for drinking. Hence, D ends up choosing not to drink as a consequence of a process of rational deliberation that has been distorted by C's exaggerating drinking's dangers. And unlike when C intercedes in D's power of recognition, here C trusts D enough to allow alcohol consumption to be among the minimally choiceworthy ends D recognizes. However, C judges that D is not as competent as she is in the appraisal of alcohol's merits. C thus allows D's rational deliberation to take drinking as one of D's *potential* ends. In this respect, C's interceding in D's power of discrimination is riskier (from her point of view) than her interceding in D's power of recognition. But should C's paternalistic intercession succeed, then D's deliberation terminates in his rejecting drinking as an end. Since here C puts more trust in D's competency as a rational will or agent, C's beliefs reflect a *moderate* mistrust of the competency of D's rational will or agency.

© Cholbi 2016

Finally, when C intercedes in D's power of satisfaction by deceiving D regarding the invitation to the watering hole, she continues to show mistrust in D's rational agency. C endorses D's end of refraining from alcohol. To this extent, C shows greater respect for D's competency in exercising his rational will than she does when she intercedes in his powers of recognition or satisfaction. In C's estimation, D can be trusted *to deliberate*, on the basis of a factually accurate understanding of the merits of alcohol consumption, about whether to drink alcohol. All the same, C judges that D is not likely to exercise his power of satisfaction adequately. Thus C does not trust D *to act* on those deliberations, since C believes D's power to realize his chosen ends is faulty. C's beliefs thus reflect a *modest* mistrust of D's rational will or agency.

The forms of paternalism are thus as diverse as the species of rationality needed to function as rational agents. More exactly, paternalistic intercessions in the power of recognition are more morally objectionable than paternalistic intercessions in the power of discrimination because the former show greater mistrust in the targeted agent's rational competency than do the latter, and paternalistic intercessions in the power of discrimination are more morally objectionable than paternalistic intercessions in the power of satisfaction because the former show greater mistrust in the targeted agent's rational competency than do the latter. Note that I have ascribed the wrongfulness of paternalism to the level of mistrust a paternalistic act *shows*, not to the level of mistrust a paternalistic actor *feels*. For the mistrust shown by a paternalistic act need not exactly reflect the mistrust the paternalistic actor has toward her target. As stated earlier, paternalists are likely to select their forms of intercession on instrumentalist grounds, interceding in whichever power will best enable them to achieve their beneficent aims. Thus, even when a paternalist mistrusts her target's more fundamental rational power, she may nevertheless find it reasonable to intercede in a less fundamental power. In fact, on the view I have offered here, a paternalist is, all other things equal, obligated to intercede in a lesser power even when her mistrust extends to a more fundamental one. Such a paternalistic act is more justified precisely because it does *not* express the full measure of distrust she has toward her target.

Consequently, because the rational powers one mistrusts need not coincide with the rational powers one intercedes in, a paternalistic act may not fully reflect the measure of the paternalist's mistrust in her target and the target's rational powers. We must therefore be careful not

 © Cholbi 2016

to assimilate too readily a paternalist's mistrust of her target's rational powers to the mistrust shown by her paternalistic act. The two avenues by which the tripartite structure of rational agency can explain paternalism's wrongfulness, one appealing to how these powers contribute to our identities as distinct rational agents, the other appealing to the levels of mistrust manifest in interceding in the different powers constituting that agency, are nevertheless closely linked. Judgments of trustworthiness are characterological. Paternalism involves the judgment that with respect to some matter with which another agent is concerned, that agent cannot be fully trusted. The paternalist's judgment that paternalism is warranted thus embeds a judgment about the agent targeted for paternalistic intercession. What I entrust you with is inseparable from my beliefs about who you are.

6. Trust, respect, and deference

A fundamental objection to my rational powers account is that the attitudes that motivate paternalistic intercessions can be epistemically warranted. Suppose that I am correct that paternalistic acts are wrong to the extent that they show mistrust in their targets' rational powers. Often though, we have evidence that others' rational powers *are* faulty—that it would be unwise for anyone, including those we might subject to paternalism, to place their trust in those powers. Self-described advocates of paternalism often appeal to this very fallibility in our powers of rational choice (Conly 2013). Why, they might ask, should we refrain from interceding in others' rational powers, given that the wrongfulness of so doing rests on its mistrust of those powers, if in fact those powers are not trustworthy? Opponents of my rational powers account may conclude that it thus requires a kind of dishonesty or make believe.

This objection can also be cast in terms of how paternalistic acts colonize or overtake our rational powers: To the extent that our identities as individuals are tied to our capacities as rational agents, we have reasons to object to being treated paternalistically because the paternalist supplants our rational agency with hers. However, it does not seem as objectionable for the paternalist to supplant our rational agency with hers if our rational agency malfunctions to such a degree that our ensuing choices and actions are not well-grounded. If our rational agency is of little instrumental value to us, its being supplanted by another's hardly counts as a loss to us.

 © Cholbi 2016

Answering this objection requires an account of (1) exactly how refraining from paternalism values rational agency properly, and (2) why the failure to value such agency is a wrongmaking feature of paternalism even when would-be paternalists have reason to judge that their targets' rational agency is not fully trustworthy.

I grant that my rational powers account rests on an *ideal* of rational agency in two senses. First, we are not *entirely* rational. Our nature as human individuals contains other elements besides our rational agency, and not everything about us that is worthy of concern is found within that rational agency. Hence, the rational powers account idealizes inasmuch as it abstracts away from these other elements in order to highlight the special role that rational agency has in explaining the moral objectionability of paternalism. Second, we are not *necessarily* rational. Not all of our choices or decisions are made on rational grounds, and some of those that are made on such grounds are not made on good rational grounds. Rational agency is thus an ideal that we value highly despite its being only imperfectly represented in our choices and actions.

But in what regard does refraining from paternalism properly value others' rational agency? 'Value' is a verb as well as a noun, an attitude directed at objects for which that attitude is apt. In the case of rational agency, its proper valuation seems to consist in honouring boundaries between agents (even boundaries rendered more porous or pliable by particular kinds of relations among agents, such as parents and children, spouses, etc., or by professional role obligations). Paternalism, as we have seen, elides such boundaries, as one agent substitutes her judgment for another's, thereby supplanting the latter's rational powers with her own. To honour such boundaries is to show *respect* for what resides on the other side of the relevant boundary from oneself (Darwall 2006, p. 268). And as Kant saw, respect for others is an attitude marking moral distance or distinctness. Respect asks us 'not to encroach upon what belongs to anyone' (Kant 1797, p. 569), including, presumably, others' rational agency and the ends engaged by that agency.

Each of us has ends for which we are concerned. Others may not share our ends, of course. They may think, without inconsistency, that our ends rest on reasons that do not strike them as forcefully as they strike us, so that what is good enough reason for us to adopt a particular end may not be good enough reason for them to do the same. Moreover, others may refrain from meddling in our ends and our pursuit thereof for reasons unrelated to respect for us as fellow

 © Cholbi 2016

agents. They may, for instance, simply be indifferent to what ends we pursue. Or they may fear the consequences of such meddling. Yet in expecting their respect, we are, in part, asking others to adopt a particular stance on those ends, namely, to *suppose* that those ends rest on adequate reasons. So in asking others to respect our ends, we are asking them to adopt an egalitarian stance akin to Kant's Kingdom of Ends: to see our ends and our pursuit thereof as meriting respect because they are manifestations of our rational agency, a respect that others do not show when they intercede in that agency in order either to better realize our ends or to realize ends they believe are better for us to pursue.

But in showing respect toward our cares and concerns as manifestations of our rational agency, others show respect toward that agency. If our ends reflect adequate reasons, then the powers that constitute our rational agency—the powers that enable practical deliberation—are likely to be operating as they should. That the reasons for our ends are adequate typically means that we are competent rational agents. Of course, that the reasons underlying our ends are adequate is consistent with our sometimes erring in our practical deliberations, in which case the presumed adequacy of our reasons is coincidental. Yet even then, respect for our rational agency militates against paternalism. Our rational powers are normative, rather than purely causal, powers, and so it is a constitutive fact about them that they can, at least in principle, be exercised inadequately. As we have observed, 'respect' carries a connotation of treating something as distinct and independent, as having boundaries that are not to be violated. Respect thus asks us to see the value of a thing as residing in its integrity or wholeness rather than its individual manifestations. As Isaiah Berlin pointed out, it would be a 'monstrous impersonation' of respect for rational agency to suppose that it merits respect only when it operates impeccably (2002, p. 180).

In refraining from paternalism, we treat another's will as authoritative for no other reason than that it is her will.[13] Hence, showing *deference* toward rational agents is an essential element of respecting rational agency, of seeing rational agency as establishing boundaries we have strong moral reasons not to trespass against. Such deference generates a high bar for the evidence needed to persuade us that someone's rational agency is faulty and thus helps to explain the

[13] As Groll (2014, p. 702) puts it, when a person is competent, we are to treat her will as 'structurally decisive', that is, as a source of de jure reasons that supersedes considerations such as her welfare, etc.

 © Cholbi 2016

general moral presumption against paternalistic intercessions in others' agency. Notice that this does not mean that whether the target of potential paternalism is trustworthy in their practical deliberations is irrelevant to paternalism or its justification. Ian Carter has recently argued that respect:

> is a substantive moral attitude that involves abstaining from looking behind the exteriors people present to us as moral agents. More precisely, while we may see behind these exteriors (for to do so is often unavoidable), if and when we do perceive people's varying agential capacities we refuse to let such perceptions count as among the reasons motivating our treatment of those people. In other words, we avoid evaluating people's agential capacities as an aid to deliberation about alternative courses of action. (2011, p. 551)

But such abstinence has a limit. At some point, evidence of another's lack of trustworthiness can raise questions about whether they are even minimally competent rational agents, in which case we descend from ideal into non-ideal theorizing about paternalism. Carter again:

> Treating persons as wholly opaque, in the sense of completely ignoring their agential capacities, would be too strong, for it would preclude those assessments that are necessary in order to have the reasonable belief that they have any agential capacities at all. We perceive individuals as moral agents because we perceive them as having at least a certain minimum of agential capacities. (2011, p. 552)

The moral demand Carter dubs 'opacity respect' thus asks us to prescind from judgments about others' rational capacities so long as their capacities exceed a minimal threshold. When they do not reach this threshold, respecting rational agency gives way to valuing rational agency, i.e., seeking to promote or restore it through means that might otherwise be objectionably paternalistic (Christman 2014, p. 373).

Respect for others' rational agency is thus a moral demand that is imperfectly sensitive to our evidence concerning others' rational agency. It asks us to discount or bracket evidence we may have concerning the competence of others' rational agency in order to respect them as rational agents. It may seem odd that a moral demand can have implications regarding our epistemic obligations—that a moral stance we take toward others can imply an epistemic stance we must take toward others. But there are instances in which our moral obligations require us to *forego* the pursuit of particular knowledge or information, such as when the right to privacy entails an obligation

 © Cholbi 2016

not to pursue certain evidence regarding another's behaviour or when the right to provide informed consent to medical interventions constrains what evidence physicians may gather in their research. The deference involved in respecting rational agency generates obligations of a different epistemic kind, compelling us (at least to some degree) to *disregard* evidence concerning an agent's rational competence. It thus operates somewhat like legal obligations, imposed on judges, to set aside evidence resting on hearsay or acquired via coercion.

In conflicts between moral normativity and epistemic normativity, the former often wins out (Wrenn 2004). In the case of paternalism, the epistemic imperfections of our rational agency do not, as some enthusiasts for paternalism would have it (Conly 2013, p. 40), mean that rational agency cannot be respected in general or when exercised inadequately. The possibility of error is built into any normative power, so that respect for rational agency requires respect for a fallible power. So our rational agency, and the choices and actions that flow from it, can and ought to be respected even on those occasions when our exercises thereof are less than sterling.

Of course, precise judgments regarding whether an individual is essentially rational but exercising her rational powers badly, as opposed to not satisfying the conditions of rational agency at all, may be difficult to come by (Cholbi 2011). A modification to my earlier example serves to illustrate this point: As far as paternalism goes, C faces a rather different set of moral challenges if, instead of trying to prevent D from picking up a drinking habit in the first place, C is attempting to stop D from drinking when D is already an alcoholic or 'problem drinker'. Should C be concerned about paternalism in the way I have represented it—as an intercession in D's rational powers—or should C conceptualize her interventions as non-paternalistic, on the grounds that while she aims to improve D's well-being, D's addiction has obliterated the rational powers intercession in which would count as paternalistic in the first place? 'Paternalistic' interventions in the choices and actions of the mentally ill raise similar worries.

Obviously, much will depend on the details of particular cases like these: how compromised the rational powers are, which powers are compromised, etc. However, my taxonomy of the rational powers can still illuminate these matters. To the extent that addicts or the mentally ill retain their rational powers, interceding in recognition is more objectionable than discrimination, which in turn is more objectionable than interceding in satisfaction, etc. And even if it is determined that the target lacks these rational powers and is therefore not a

 © Cholbi 2016

rational agent, a moral justification for interceding is still needed despite these intercessions not rightfully being classified as paternalistic (Feinberg 1986, pp. 12-3). That an intercession in others' actions or choices is not paternalistic does not make it morally unproblematic. Such a justification may rest on the value of rational agency, but not on respecting it. The justification may instead rest on aims such as *protecting* or *promoting* rational agency. My aim here has been to show how the wrongfulness of various paternalistic acts can be mapped onto the architecture of rational agency, not to appraise possible justifications for intervening in the choices or actions of beings who are not rational agents. That 'non-ideal' project must be addressed elsewhere (Jaworska 1999; Schapiro 1999; Cholbi 2002).

7. The varieties of paternalism

Finally, let us note that the rational powers account helps make sense of some widely held views about the moral objectionability of different varieties of paternalism.

For instance, strong paternalism faces a higher burden of proof than does weak paternalism. Strong paternalism is directed toward a person whose choice of ends is, in the estimation of the paternalistic intervenor, mistaken or irrational. Weak paternalism, in contrast, is directed toward a person whose choices and actions will, in the estimation of the paternalistic intervenor, frustrate or undermine that person's ends. In general, strong paternalism seems more morally problematic than weak paternalism for reasons that my rational powers account illuminates. The strong paternalist mistrusts her target's capacity to select worthwhile or choiceworthy ends. Hence, the strong paternalist is likely to be more willing to intercede in those powers—the powers of recognition and discrimination—intercession in which shapes the ends a person chooses. Again, this willingness need not translate directly into interceding in these more fundamental powers. But the strong paternalist, given her willingness to shape both her target's ends and the chosen means to those ends, would presumably not hesitate to intercede in the powers associated with identifying and choosing ends (recognition and discrimination).

Weak paternalism, on the other hand, is motivated by concern not for the 'wisdom, prudence, or dangerousness' *per se* of its target's chosen ends but by concern that the target's choices and actions are, as means to her chosen ends, at odds with those ends (Feinberg

 © Cholbi 2016

1986, p. 12). The weak paternalist thus aims to make her target's choice of action align with the target's ends. The weak paternalist is thus likely not to try to modify her target's ends but to intercede in the target's power of satisfaction so that the target's choices in fact satisfy her ends. Hence, while not all instances of weak paternalism will be less morally objectionable, either *qua* paternalistic or from an all-things-considered perspective, than all instances of strong paternalism, the view I have developed thus far explains why this will usually be the case: Strong paternalism intercedes in rational powers the intercession in which shows greater mistrust in its target's rational competency than similar weak paternalist intercessions in the power of satisfaction.

Likewise, my version of the rational will view sheds light on many philosophers' relative enthusiasm for soft over hard paternalism. Soft paternalists maintain that while we may not act paternalistically to prevent a person from knowingly and voluntarily engaging in self-harm, we may act paternalistically if doing so is necessary to determine whether a person is knowingly and voluntarily intending to self-harm. To use a timeworn example, soft paternalists do not deny that someone has the right to walk across a damaged bridge, but they permit others to interfere with her intention to do so if they have reason to believe that she is not aware of its fragile condition. Hard paternalists reject this condition, allowing paternalism to prevent self-harm even when a person knowingly and voluntarily intends such harm to herself. Hard paternalists thus assert that we may sometimes prevent others from acting on ends they have rationally chosen. Hard paternalism is thus more morally objectionable than soft paternalism, according to my version of the rational will view, because it intercedes in a more fundamental rational power—the power of discrimination—than does soft paternalism, which is aimed at ensuring that agents are exercising their power of satisfaction in accordance with their rationally chosen ends.

8. Conclusion

Gerald Dworkin has observed that debates about paternalism, while 'about facts … and how effective various policies are', are also 'crucially about different ideals of the person' (2013b). Here I have argued for a version of the rational will view of paternalism—the rational powers account—that takes very seriously the ideal of the person as a rational agent, with particular attention to the architecture of

 © Cholbi 2016

rational agency and the ways in which various paternalistic acts or policies can intercede in that architecture.

I have not taken up some disputes regarding paternalism. For example, I have not weighed in on whether paternalism intercedes in another's rational will so as to advance her interests or well-being or whether paternalism can also occur when one intercedes in another's rational will for the sake of her ends, including her non-self-interested or other-regarding ends. Nevertheless, by explaining a wide range of ethical data about paternalism in a parsimonious way, my rational powers account of paternalism appears to offer the foundation of a plausible philosophical theory of paternalism. It provides powerful explanations of why rational suasion usually deflects charges of paternalism; why consenting to intercessions in one's rational agency negates paternalism; why we ordinarily believe that strong paternalism is more objectionable than weak paternalism; and why we ordinarily believe that hard paternalism is more objectionable than soft paternalism. And although my defence of the rational powers account does not strive to provide a complete casuistry of paternalistic acts and policies, it suggests how such an account might be developed so as to make sense of the intuitively plausible claim that paternalism is presumptively, but not necessarily, impermissible. Admittedly, the rational will view's characterization of paternalism is not normatively *neutral*, since it refers to wrong or badmaking features of paternalism (Bullock 2015). However, it is also not objectionably *partial*, since it also refers to apparently goodmaking features of paternalism, namely, that (successful) paternalism helps its targets secure their ends or the ends they ought to seek. Hence, the rational will view does not illicitly close off debates about the pros and cons of particular instances of paternalism. After all, proponents of paternalism can grant that such normatively 'loaded' definitions are correct while attempting to show that the bad or wrongmaking feature is either merely *prima facie* or is outweighed by its goodmaking feature.[14]

References

Anderson, Elizabeth 1999: 'What is the Point of Equality?' *Ethics* 109, pp. 287–337.

[14] I gratefully acknowledge the feedback and insights that David Adams, Emma Bullock, Nico Cornell, Carl Cranor, John Davis, Johannes Drerup, Megs Gendreau, Kalle Grill, Andres Moles, Steven Munzer, Chris Naticchia, Maura Priest, Jonathan Quong, Peter Ross, Peter Schaber, Danny Scoccia, Tom Walker, and three referees for this journal provided on earlier versions of this manuscript.

 © Cholbi 2016

Berlin, Isaiah 2002: *Four Essays on Liberty*. Oxford: Oxford University Press.

Bullock, Emma C. 2015: 'A Normatively Neutral Definition of Paternalism'. *Philosophical Quarterly* 5, pp. 1–21.

Carter, Ian 2011: 'Respect and the Basis of Equality'. *Ethics* 121, pp. 538–71.

Cholbi, Michael 2002: 'Suicide Intervention and Non-Ideal Kantian Theory'. *Journal of Applied Philosophy* 19, pp. 245–59.

——2011: 'Kantian Paternalism and Suicide Intervention'. In Coons and Weber 2013, pp 115–33.

Christman, John 2014: 'Relational Autonomy and the Social Dynamics of Paternalism'. *Ethical Theory and Moral Practice* 17, pp. 369–82.

Conly, Sarah 2013: *Against Autonomy*. Cambridge: Cambridge University Press.

Coons, C., and M.E. Weber, (eds.) 2013: *Paternalism: Theory and Practice*. Cambridge: Cambridge University Press.

Cornell, Nicolas 2015: 'A Third Theory of Paternalism'. *Michigan Law Review* 113, pp. 1–44.

Darwall, Stephen 2006: 'The Value of Autonomy and Autonomy of the Will'. *Ethics* 116, pp. 263–84.

De Marneffe, Peter 2006: 'Avoiding Paternalism'. *Philosophy and Public Affairs* 34, pp. 68–94.

Dworkin, Gerald 2013a: 'Defining Paternalism'. In Coons and Weber 2013, pp. 25–38.

——2013b: Review of S. Conly, Against Autonomy. Notre Dame Philosophical Reviews (http://ndpr.nd.edu/news/40382-against-autonomy-justifying-coercive-paternalism/), 4 June 2013.

Egonsson, D., J. Josefsson, ,B. Petersson, ,T. Ronnow-Rasmussen, and I. Persson, (eds) 2001: *Exploring Practical Philosophy: Essays in Honour of Ingmar Persson*. London: Ashgate.

Feinberg, Joel 1986: *Harm to Self*. Oxford: Oxford University Press.

Gregor, M. (ed and trans) 1996: *Practical Philosophy (Cambridge Edition of the Works of Immanuel Kant)*. Cambridge: Cambridge University Press.

Grill, Kalle 2007: 'The Normative Core of Paternalism'. *Res Publica* 13, pp. 441–58.

Groll, Daniel 2012: 'Paternalism, Respect, and the Will'. *Ethics* 122, pp. 692–720.

 © Cholbi 2016

Jaworksa, Agnieszka 1999: 'Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value'. *Philosophy and Public Affairs* 28, pp. 105–138.

Kant, Immanuel 1797: *Metaphysics of Morals.* In Gregor 1996, pp. 363–603.

Kleinig, John 1983: *Paternalism.* Totowa, New Jersey: Rowman and Allenheld.

Korsgaard, Christine 1996: *The Sources of Normativity.* Cambridge: Cambridge University Press.

Mill, J.S. 1859: *On Liberty.*

Parfit, Derek 2001: 'Rationality and Reasons'. In Egonsson, Josefsson, Petersson, Ronnow-Rasmussen, and Persson 2002, pp. 17–39.

Quinn, Warren 1994: 'Putting Rationality in Its Place'. In his *Morality and Action* (Cambridge: Cambridge University Press), pp. 228–51.

Quong, Jonathan 2011: *Liberalism Without Perfection.* Oxford: Oxford University Press.

Rawls, John 2005: *Political Liberalism*, expanded second edition. New York: Columbia University Press.

Schapiro, Tamar 1999: 'What is a Child?' *Ethics* 109, pp. 715–738.

Shiffrin, Seana 2000: 'Paternalism, Unconscionability Doctrine, and Accommodation'. *Philosophy and Public Affairs* 29, pp. 205–50.

Tsai, George 2014: 'Rational Persuasion as Paternalism'. *Philosophy and Public Affairs* 42, pp. 78–112.

Wrenn, Chase 2004: 'Hypothetical and Categorical Epistemic Normativity'. *Southern Journal of Philosophy* 42, pp. 273–90.

© Cholbi 2016

Metanormative Contextualism and Normative Uncertainty

JOHN PITTARD
Yale University
john.pittard@yale.edu

ALEX WORSNIP
University of North Carolina at Chapel Hill
aworsnip@unc.edu

We offer a new argument in favour of metanormative contextualism, the thesis that the semantic value of a normative 'ought' claim of the form '*S* ought to Φ' depends on the value of one or more parameters whose values vary in a way that is determined by the context of utterance. The debate over this contextualist thesis has centred on cases that involve 'ought' claims made in the face of uncertainty regarding certain descriptive facts. Contextualists, relativists, and invariantists all have plausible ways of explaining these cases, and one could reasonably judge the debate between these views to be a stalemate. We argue that this stalemate can be broken by shifting focus to a case that involves normative uncertainty rather than descriptive uncertainty. While relativist and invariantist rivals of contextualism can give plausible accounts of the descriptive uncertainty cases, only contextualism can provide a plausible account of the normative uncertainty case.

Introduction

Jill says she ought to do action *A*, but Jack says that Jill ought to do some incompatible action *B* instead. These sound like inconsistent normative claims. But in some cases of this sort, cases where Jill is uncertain about something that Jack is not uncertain about, it seems that Jill and Jack both speak truly. The puzzle presented by such cases has recently received a number of competing solutions: relativist solutions, contextualist solutions, and non-relativist 'invariantist' solutions. Each of these approaches has resourceful defenders, and at present one could reasonably judge the debate to be a stalemate. In this paper, we argue that this stalemate can be broken if we focus on a type of case that differs in an important respect from the type of case that has been at the centre of the discussion thus far: specifically, on a case involving *normative* uncertainty rather than *descriptive* uncertainty. Focusing on such a case reveals the inadequacy of the relativist

Mind, Vol. 126 . 501 . January 2017 **© Pittard and Worsnip 2016**
doi:10.1093/mind/fzv208 Advance Access publication 19 October 2016

and non-relativist invariantist proposals and allows one to mount a clear and compelling argument in favour of metanormative contextualism.

Normative uncertainty—and how one ought to respond to it—is in its own right the subject of lively current debate in normative theory. Yet cases of normative, as opposed to descriptive, uncertainty have generally been passed over in the literature on contextualism. This is ironic since, as we will argue, such cases of normative uncertainty actually provide the key to securing the case for contextualism.

For the purposes of this paper, we stay largely neutral on how the details of a contextualist semantics for normative terms are best developed, and silent on how to address the various objections that have been levelled at contextualism. If our argument is sound, then some form of contextualism has to be correct; the details are challenges for future work.

1. The original descriptive uncertainty cases

Let us begin with a representative example of the sort of case of descriptive uncertainty that has been the focus of the existing debate. Consider the following case, adapted from Jackson (1991, pp. 462-3):[1]

> Sarah is an herbalist in training. While collecting herbal treatments in the woods, she encounters Sam, who has been bitten by a poisonous snake and who will die soon if not treated immediately. Sarah has three herbs at her disposal: herb A, herb B, and herb C. She knows that herb C will save Sam's life, but will leave him with mild hypersensitivity to the sun. She also knows that one of herbs A and B will completely cure Sam with no side effects, but that the other will kill him. Unfortunately, she doesn't remember which is which. 'Since I can't remember which of herbs A and B is the one that will cure Sam and which is the one that will kill him, I ought to give him a treatment of herb C', she says to herself. The next day, she describes the episode to her herbology professor, who sighs and says, 'You ought to have administered only herb B. Now Sam will need to wear sunscreen wherever he goes'.

Sarah claims that she ought to administer herb C. Her professor claims that she ought to have administered only herb B. Because these

[1] For similar examples, see Regan (1980, p. 265, n. 1); Parfit (2011, 1: p. 159).

 © Pittard and Worsnip 2016

are incompatible courses of action, the claims of Sarah and her professor appear on face to be incompatible.[2] Yet it seems that Sarah and her professor both speak truly when making these normative claims.

Metanormative contextualism offers a solution to this puzzle. According to metanormative contextualism (from here on, just 'contextualism'), the semantic value of a normative 'ought' claim of the form '*S* ought to ϕ' depends on the value of one or more parameters whose values vary in a way that is determined by the context of utterance.[3,4] The proposal is, then, that 'ought' has one semantic value in Sarah's mouth, and another semantic value in her professor's mouth. Hence, both claims can be true.

There is a range of competing contextualist proposals about how to make good on this general idea. On one proposal, there is one parameter for standards and a second parameter for information, so that the meaning of '*S* ought to ϕ' depends on which standards and information set are contextually relevant.[5] Given this semantics, one way of interpreting the present example is to say that Sarah's statement amounts to a claim about what action has the highest expected value in light of the information available to *her*, while the professor's statement amounts to a claim about what action has the highest expected

[2] Like others who have discussed cases like Sarah's, we take it for granted that Sarah's situation is *not* a deontic dilemma where it is true (from a single context of assessment and without any variation in the semantic value of 'ought') *both* that Sarah ought to give herb C *and* that she ought to give herb B alone. When we come to the case of normative uncertainty later, we will return to the possibility of deontic dilemmas (§3.3).

[3] There are also non-normative usages of 'ought', such as the 'evaluative "ought"', and the '"ought" of expectation'. We think that the most plausible semantics will give a unified account of all normative usages of 'ought'—both concerning how one ought to act and concerning what one ought to believe—and probably of non-normative usages too. However, here we focus on normative usages about how one ought to act. The contextualist semantics should also extend unproblematically to normative claims featuring other related terms (e.g., 'should' and 'may').

[4] A contextualist semantics for 'ought' and related terms has recently been defended by, among others, Björnsson and Finlay (2010); Brogaard (2008); Cariani, Kaufmann, and Kaufmann (2013); Dowell (2012 and 2013); Finlay (2014); Henning (2014); and Wedgwood (2006 and 2016). Snedegar (2013) and Cariani (2013) defend a contrastivist semantics for 'ought', which for our purposes is a form of contextualism; their reasons for preferring the contrastivist variant do not bear on the kind of cases that will be the focus of our discussion. Strangely, while contextualism is non-standard amongst metaethicists, and explicit defences of it in metaethics are mostly very recent, contextualism about modals such as 'ought' (including their deontic uses) has long been orthodox in philosophy of language, following the semantics of Kratzer (1981), which several of the above accounts draw and build upon. One forerunner of contextualism in the metaethical literature, however, is Dreier's (1990) 'speaker relativism'.

[5] See, e.g., Björnsson and Finlay (2010).

 © Pittard and Worsnip 2016

value given the information available to *him*. (Alternatively, the relevant standards might vary across the two statements if the professor's statement amounts to a claim about which action has the highest *actual*—rather than expected—value.) To stave off potential misunderstanding, it is worth emphasizing that the contextually-relevant set of information need not be the information that is accessible to or known by the speaker in every case. The speaker's interests and intentions could render some other information set contextually relevant.[6]

There are reasons for thinking that this simple proposal is not ultimately adequate.[7] But we do not aim to take a stand here on the question of how exactly to develop the contextualist semantics. What all versions of contextualism have in common is that they posit different semantic values for different utterances of 'ought', in a way that is not explained by brute ambiguity but rather by a unified semantics

[6] Cf., Dowell's (2013) 'flexible contextualism'; see also Kolodny and MacFarlane (2010, p. 122).

[7] In particular, a catch-all notion of 'information' obscures two very distinct effects of information changes, which may be best modelled separately in the semantics. To illustrate this, consider the orthodox Kratzer (1981) semantics, which includes two parameters: a 'modal base' and an 'ordering source'. The modal base consists of a set of propositions, which restrict the worlds over which a modal like 'ought' quantifies (since the modal quantifies only over the worlds in which the propositions in the modal base are true). The ordering source consists of a set of norms, which rank these worlds by how well they satisfy the norms in question. So '*S* ought to ϕ' is true iff *S* ϕ's in all the highest-ranked worlds. Now, the difference in information between Sarah and her professor is best understood as resulting in a shift in the contextually relevant ordering source (cf., Dowell 2012). For Sarah's claim, the ordering source is expected value given *her* information, while for the professor the ordering source would be expected value given *his* information (or perhaps actual value). Although in one sense the 'standards'—expected value—have remained fixed, one gets a different ordering of worlds depending on whose information is relevant in calculating such expected value.

On the other hand, there is a different category of information shifts that are best understood as taking effect on the modal base. Such cases involve information that bears on what options are possible or 'live' for the subject. Such information may include external constraints on the subject's choice, but can also in certain contexts include the subject's plans, intentions, and future actions that, when held fixed, effectively rule out potential courses of action and thus restrict the range of 'live' options. For example, if Sheila is giving away $1,000 to a charity and has narrowed her options down to charity A and charity B, Jim could truly say 'You ought to give the money to charity B' even if, on the standards that constitute the ordering source for Jim's utterance, giving $1,000 to charity C would be better than either A or B. This is because when we add the proposition that Sheila will not give to charity C to the modal base, we are restricted to worlds in which she does not give to charity C, and the best *remaining* worlds, given the same ordering source, might be those in which she gives to charity B. Depending on Jim's purposes in speaking, he may wish either to restrict himself to the worlds in which Sheila does not give to charity C to express this 'ought of the second-best', or to talk about what Sheila *ideally* ought to do, not holding fixed her current plans and intentions.

In this paper, the first kind of information-sensitivity—that which takes effect on the ordering source—will generally be the one of relevance.

 © Pittard and Worsnip 2016

with one or more contextually-determined parameters. Though this leaves a lot open, it is a distinctive enough claim for our purposes in this paper, and our argument is pitched at a level of generality that favours this general proposal over other competing proposals. We now turn to an explication of these competing proposals for dealing with cases of descriptive uncertainty like the one we have just given. This discussion will help make clear why our argument for metanormative contextualism, which appeals to cases of *normative* uncertainty, is stronger than arguments relying on cases like the one above.

2. Non-contextualist proposals for resolving the cases

2.1 *Soft invariantism*

The 'invariantist', as we will use the term, denies both that 'ought' is context-sensitive, as the contextualist holds, and that the truth of 'ought' claims is assessment-sensitive, as the relativist holds (see §2.3 below). Although this view may not be a popular one in semantics, we wish to stress that some kind of invariantism is still the dominant assumption in ethics and epistemology.[8] So it is important for us to engage it in this paper.

Let us start by distinguishing two kinds of invariantist view. According to the *soft invariantist*, both Sarah and her professor speak truly, but this is because they use 'ought' in different *senses* rather than because 'ought' is context-sensitive; uses of 'ought' may sometimes be ambiguous between these different senses.[9] The *hard invariantist*, on the other hand, claims that Sarah's claim really is inconsistent with her professor's claim, and that at least one of them speaks falsely, strictly speaking, though s/he may implicate something true in the course of so doing.

We begin with soft invariantism, according to which Sarah and her professor use 'ought' in different senses. An obvious place to start is with the common suggestion that 'ought' has an 'objective' sense and

[8] Many writers in the ethics and epistemology literatures assume that 'ought' is ambiguous between two 'senses': objective and subjective. See, e.g., the many references collected in Feldman (1988), including Brandt, Pollock, Goldman, Unger, Alston, Kornblith, Moser, and Kvanvig, as well as, more recently, e.g., Setiya (2004), Gibbard (2005), and Schroeder (2009). This is a 'soft invariantist' view, as characterized below, in contrast to a genuine contextualist view.

[9] Ambiguity and context-sensitivity should not be confused. An ambiguous term has multiple meanings, each of which needs its own semantics. A context-sensitive term, by contrast, has a single semantics, with one or more parameters which get supplied by the context.

 © Pittard and Worsnip 2016

a 'subjective' sense. Perhaps Sarah uses the subjective 'ought', which concerns what she ought to do given her subjective position, and the professor uses the objective 'ought', which concerns what she ought to do from the point of view of all the facts.

Clearly this proposal has some affinities with the contextualist idea. However, the idea that there are just *two* 'ought's—objective and subjective—is subject to a serious, and in our view fatal, problem of regress.[10] Suppose that we posit just two 'ought's: for example, one that pertains to what you ought to do given all the facts and another that pertains to what you ought to do given your doxastic attitudes (such as your credences and beliefs).[11] The latter 'ought' helps us to make sense of claims about what you ought to do *given your uncertainty about the facts*. But these two 'ought's are not enough to help us make sense of claims about what one ought to do given further layers of uncertainty. Perhaps, for example, what you ought to do *given your uncertainty about the facts* is to perform the action with maximal subjective expected utility. But clearly, you can be (justifiably) uncertain about which action satisfies this criterion, and sometimes you should not take the steps that would resolve all of this uncertainty. For example, circumstances might demand that you act quickly, and determining which action has the maximal subjective expected utility might involve lengthy expected utility calculations that cannot be completed before the moment of decision.

Naturally, we might want to speak about what you ought to do given this deeper level of uncertainty. We might say, for instance, that you ought to perform some action *A* even if you are certain that it is not the action with the highest subjective expected utility as long as (i) you believe that the expected utility of *A* is high, (ii) you do not believe of some alternative action that it will have a higher expected utility, and (iii) there are no alternative actions that are obviously worthy of consideration and that you would (upon initial consideration) believe to have a higher expected utility than *A*. Accounting for this claim requires that we posit an additional semantic value for 'ought'; to accommodate this, the advocate of two senses of

[10] For related discussions, see Sepielli (2014); Kagan (ms.); Wells (ms.); and MacAskill (ms.).

[11] If it's already occurring to you that there must be other 'ought's—like what you ought to do given your *evidence*—then this only serves to further illustrate our point. See Feldman (1988) for a nice overview of how the objective/subjective distinction gets drawn in various non-equivalent ways by different writers, such that the literature as a whole is dealing in more than two notions.

 © Pittard and Worsnip 2016

'ought'—objective and subjective—will either have to admit that at least *one* of those 'ought's is context-sensitive, or posit a third sense of 'ought'.

But that's not it: the problem will then iterate once more for that third 'ought'. For starters, whether an alternative action is 'obviously' worthy of consideration (in the way that clause (iii) refers to in the characterization of the third 'ought') itself depends on whether obviousness is understood in a more objective or more subjective way. For some purposes we might be interested in speaking about what you ought to do given a more objective construal of obviousness, and for other purposes we might want to speak of what you ought to do given what was and was not obvious *to you*. Additionally, one can sometimes be uncertain about the content of one's beliefs concerning the comparative expected utility of different actions, and we might want to speak about what you ought to do *given that you are uncertain whether you believe that the expected utility of* A *is high* (for example). Making sense of claims about what you ought to do given further layers of uncertainty (or given what does and does not obviously occur to you) will require positing further semantic values for 'ought'. One may accommodate this by positing additional senses, but the problem will reoccur for these additional senses, multiplying senses of 'ought' *ad infinitum*. It is not plausible that 'ought' is *ambiguous* between such countless readings. Contextualism avoids the need to posit such ambiguity: it posits one 'ought', with one or more contextually-sensitive parameters. Ultimately, then, the best version of the objective and subjective multiple senses proposal collapses into contextualism, on which there is a range of 'ought's, some more subjective and some more objective (and perhaps on different dimensions of objectivity).

But there are alternative proposals for multiple senses of 'ought' that the soft invariantist can appeal to, and that are less easily dealt with. For example, one might claim that Sarah is using 'ought' in a moral sense, while her professor is using 'ought' in an *evaluative*, non-moral sense.[12] When I exclaim that one ought to be able to take a walk in the park without getting eaten up by mosquitoes, it seems that I am not making a claim about requirements (moral or otherwise) that pertain to any agent, but am rather making a claim about what states of affairs would be *best*. Similarly, perhaps the professor's statement should be understood as a claim about what would have been

[12] See Schroeder (2010) for a view that might inspire this kind of proposal.

 © Pittard and Worsnip 2016

best rather than as a claim about what reasons Sarah had or what morality or rationality required of her.

Now, it seems to us that even this kind of soft invariantist will be forced to concede that at least one of the senses of 'ought' at play will require a contextualist semantics. Suppose that ingesting a certain amount of soil can also prevent the ill effects of snake venom, and that *unlike* herb B, soil does *not* have the property of neutralizing some of the mild positive health benefits that snake venom is known to have. So in an ideal scenario, snake bite victims would not be given herbs but would instead be given the right amount of soil. But suppose further that the amount of soil administered must be within a *very* narrow range to save the victim, a range whose values depend on the precise amount of venom in the blood, the patient's weight, the alkalinity of the soil, and many other factors that could not be reliably measured without specialized lab equipment, equipment that is *not* available in the woods. Ingesting a quantity of soil that is not in this narrow range will result in the certain death of the victim. If the professor knows all this, it still seems that he can truly affirm that Sarah ought to have administered herb B, even though he knows that there was some amount of dirt that, if administered by Sarah, would have resulted in a better outcome.

The soft invariantist response to the original case needs to explain how, in this modified case, the professor could still reasonably say that Sarah ought to have administered herb B even though he knows there is some action that would have been better. It seems to us that the soft invariantist must concede that evaluative 'ought' claims do not always express a view about what is best absolutely, but may sometimes express a view about what is best from among some set of options determined by conversational context. So there will be pressure to concede a contextualist semantics for 'ought' claims. Of course, the type of example that we would need to pressure the soft invariantist into contextualism would be different depending on how the soft invariantist proposes to disambiguate various 'ought' claims. But no matter what senses of 'ought' are appealed to, there will probably be a modified example that can be accommodated in a plausible way only by appealing to a contextualist semantics for at least one of these senses.

However, even if we are right that the soft invariantist will be pressured into accepting a contextualist semantics for 'ought' in at least one of its senses, the soft invariantist might respond to this pressure by allowing that the *non-normative* 'ought' is

 © Pittard and Worsnip 2016

context-sensitive while denying that the normative 'ought' is context-sensitive. In this way, she could continue to appeal to multiple senses as a way of denying metanormative contextualism, which is a thesis about the normative use of 'ought' in particular.

2.2 Hard invariantism

According to the hard invariantist, there is just one sense of 'ought' in play in the kinds of cases that motivate contextualism, and it is not context-sensitive or assessment-sensitive. While this might seem like an extreme or rigid view, there are two manoeuvres available to the hard invariantist to explain away the apparent variability of 'ought', which together can seem to adequately deal with the range of ways that we might develop the cases of descriptive uncertainty such as that of Sarah.

The first is to claim, drawing on the work of Jonathan Dancy (1977; 2000, ch. 3) and especially John Broome (1999; 2004; 2007; 2013, esp. ch. 3) that 'ought' can sometimes take wide scope over a conditional.[13] Then, one can say that the following two claims are both true:

(1) Sarah ought to administer herb B (rather than herb C).

(2) Sarah ought [to make it true that] (if she is uncertain as to which of herb A and herb B is the optimal herb and which is the killer herb, then she administers herb C).

(1) and (2) are consistent, even on the assumption that the semantic value of 'ought' at issue in the two claims is identical. Moreover, we cannot derive any inconsistency from them even with the further premise that, as a matter of fact, Sarah is uncertain as to which of herb A and herb B is the optimal herb and which is the killer herb.

[13] There is a considerable literature, somewhat orthogonal to our purposes here, on whether rational requirements are best-understood as having a wide-scope logical form. (See the aforementioned Broome and Dancy references, as well as e.g., Brunero (2010; 2012), and Way (2010; 2011) for the view that they are, and Schroeder (2004), Kolodny (2005), and Lord (2011; 2014) for the view that they are not.) As has only been explicitly appreciated fairly recently (cf. Dowell 2012; Silk 2014; Worsnip 2015a), this question is independent of the question of whether any particular class of ordinary utterances are best understood as *semantically expressing* wide-scope 'ought's, let alone whether this mitigates the need for a non-invariantist semantics. While some wide-scopers about rational requirements seem to suggest this semantic view (Broome 2013; Brunero 2010; Dancy 1977), there are powerful critiques of it going back all the way to Hansson (1969); van Fraassen (1972); Lewis (1973); and Kratzer (1981), developed in the current literature in Dowell (2012); Silk (2014), and Worsnip (2015a). Still, this does not necessarily speak against the view that the fundamental requirements of rationality are wide-scope in their logical form; see Worsnip (2015a).

 © Pittard and Worsnip 2016

That is because, as a general matter, the following scheme (often known as 'factual detachment' for the wide-scope 'ought') is invalid:

A
Ought(A→B)
Therefore, Ought(B)

How does this help the hard invariantist to account for the intuition that both Sarah and her professor speak truly? It is easy to explain why the professor speaks truly: his utterance expresses (1), which is true. As for Sarah, the hard invariantist has a few different options. She can claim that Sarah's utterance actually expresses the wide-scope claim (2), despite the absence of any obvious syntactic marker that it does so. Less ambitiously, she can claim that Sarah's utterance seems true to us because it is 'in the ballpark' of a truth, namely (2).

There is an important complication, however. While there is consensus that factual detachment fails for wide-scope 'ought's, a schema known as 'deontic detachment' is classically assumed:[14]

Ought(A)
Ought(A→B)
Therefore, Ought(B)

Standard deontic logic (specifically, the 'K axiom' of standard deontic logic) vindicates deontic detachment for wide-scope 'ought's. And we can develop Sarah's case so that it is clear that she *ought* to be uncertain about which herb to prescribe. For example, if Sarah had learned that one of herbs A and B would save Sam and one would kill him, but had not yet been told which is which, then it would be the case that Sarah ought to be uncertain as to which is the killer herb and which is the optimal herb. In such a case, by deontic detachment, Sarah ought to prescribe herb C, so an inconsistency with (1) follows.

A hard line wide-scoper may simply read such a result as a reason to reject the validity of deontic detachment. Indeed, Broome denies deontic detachment.[15] We ourselves find deontic detachment plausible, both in the abstract and in the particular case of Sarah.[16] In any case, however, there is a second manoeuvre available to the hard

[14] Cf., e.g., Hare (1971); Greenspan (1975).

[15] Broome (2013, pp. 120–1, 127–8).

[16] See Worsnip (2015b, §3.8) for critical discussion of Broome's arguments against deontic detachment. Moreover, an even weaker principle than deontic detachment suffices to make trouble for the wide-scoping strategy. This weaker principle says that Ought(A→B), Ought(A) and (the actual fact that) A altogether entail Ought(B). This weakened principle is immune to Broome's counterexamples to deontic detachment.

 © Pittard and Worsnip 2016

invariantist that can deal with cases where it is clear that the agent's uncertainty is justified. Specifically, the hard invariantist can deny in such cases that the third party (in this case the professor) really speaks truly, by claiming that the third party mistakenly projects his or her own epistemic position onto the agent. The hard invariantist might suggest that once we clarify that the agent's uncertainty was justified, the tendency to endorse the third party's judgment dissipates at least somewhat.[17] Moreover, to the extent that the intuition remains, we may *ourselves* be projecting our knowledge of the stipulated fact to which the third party is supposed to be privy in the example. The hard invariantist can appeal to the fact that Sarah could reasonably protest, in response to the professor, 'That's not fair! You hadn't yet told us which herb is the cure! I did exactly what I ought to have done'.

So we have a pleasing two-prong recipe for a hard invariantist resistance of contextualism in uncertainty-based cases like that of the herbalist. If it seems that the subject *ought not* to have been uncertain, then we say that it is the professor (or analogue to such a figure) that speaks truly, and that, while the subject speaks falsely, her assertion communicates something true, namely a wide-scope requirement governing the relationship between her doxastic states and actions. If it seems that the subject *ought* to have been uncertain, then we say that it is the subject who speaks truly, and that the professor-figure engages in mistaken projection that ends in a false utterance.

Again, we think that this attractive strategy fails in the case of normative uncertainty that we develop below. But before showing why, we have one more alternative position to explicate: that of the relativist.

2.3 Relativism

Like the hard invariantist, the metanormative relativist holds that 'ought' takes on the same semantic value in the 'ought' claims of both Sarah and her professor. But the relativist seeks to more fully accommodate the intuition that both Sarah and the professor speak

[17] By analogy, consider the tendencies for post-hoc analysis of strategic decisions by sports fans. For example, roughly speaking, in baseball, attempting to steal a base raises the expected chance of the running team's scoring just if the baserunner is successful in stealing bases more than 75% of the time. Suppose a baserunner is generally successful 95% of the time; clearly, then, his decision to run is (given favourable circumstances) a good one. But, if he fails—if it turns out to be one of the unsuccessful 5% of attempts—sports fans will be liable to say, 'Argh! That was an awful decision! He shouldn't have tried to steal that base!' This seems a clear projection of what the fans know after the base-stealing attempt onto the epistemic position of the baserunner.

 © Pittard and Worsnip 2016

truly. According to the relativist, (many) normative claims are neither true nor false absolutely, but are only true or false relative to some 'context of assessment' (Kolodny and MacFarlane 2010; MacFarlane 2014). A context of assessment may include a particular body of information, as well as particular standards of assessment. It is important to see that the *same* token utterance can be assessed as true or false at various different contexts of assessment. Although individuals making assessments occupy contexts of assessment, or 'perspectives', there is no determinate 'assessor' for an individual utterance in the way that there is a determinate speaker and subject; rather, there are as many possible contexts of assessment for the utterance as there are possible sets of information and standards.

In the case of Sarah, the relativist can say the following. Assessed from the professor's vantage point, the proposition expressed by Sarah's claim that she ought to administer herb C is false. But assessed from Sarah's less informed perspective, this very same proposition is true. Some normative statements, at least, can be 'true for you, but not for me'.

To be clear, the verdict is not straightforwardly that the professor's claim is true and that Sarah's claim is true. In fact, there is no *single* context of assessment from which we can say both of those things, since on the relativist account (by contrast to the contextualist account), the professor and Sarah express inconsistent propositions. But what we can say is that both speakers say something which is true at their own context of assessment. We are not by any means required to take up the same perspective that these speakers occupy when assessing the truth of their utterances; unlike contextualism, relativism gives no special privilege to the speaker's context. However, the relativist can still explain why both claims seem sensible and true from their own contexts of utterance. This helps to explain why we feel tempted to endorse both claims as true, even though there is no single context of assessment from which we can actually do that.

Indeed, some relativists claim that relativism actually does *better* than contextualism in accounting for our intuitive reactions, since we may hesitate to say unequivocally that both Sarah and the professor speak truly, or that there is no disagreement between them. While their individual claims both seem reasonable, we may experience them as being in at least a certain amount of tension. Relativism can nicely account for this equivocal reaction.[18]

[18] See Kolodny and MacFarlane (ms.). In reply, many contextualists dispute whether cases such as that of Sarah and the professor really amount to disagreements in the strong sense in

 © Pittard and Worsnip 2016

3. A new normative uncertainty case

The cases that have thus far been the focus in the debate over metanormative contextualism have been cases involving what we will call *descriptive uncertainty* rather than *normative uncertainty*. A brief example will illustrate the distinction we have in mind. If you know that Abby ought to repay her lender, but do not know whether her lender is Beth or Carl, you may as a result be uncertain about whether or not Abby ought to pay money to Beth. While this is uncertainty about a normative claim, the uncertainty arises from uncertainty about a purely descriptive fact (the identity of Abby's lender) and is compatible with your being completely certain about all the general normative principles. If, on the other hand, you know that Beth loaned money to Abby but are uncertain about whether there is a *prima facie* obligation to repay lenders, then your uncertainty about whether Abby ought to pay money to Beth is fundamentally normative in character: your uncertainty could persist even if you were completely certain about all the purely descriptive facts of the case.

The case involving Sarah the herbalist that has been the focus here is a case where it is purely descriptive uncertainty that accounts for the utterance of multiple 'ought' claims that (superficially) appear to be inconsistent. Similarly, the most prominent example featured in recent debates is also clearly one featuring descriptive rather than normative uncertainty.[19] Moreover, the theoretical development of the basic idea that there is more than one 'ought' in play in such cases generally assumes that it is descriptive rather than normative uncertainty that generates the need to appeal to multiple senses or semantic values for 'ought'. The simple idea is that we have different possible bodies of descriptive information, but hold general normative principles fixed, as providing something like a function from that body of information to claims about what ought to be done relative to the information. So, for example, if utilitarianism is true, we just work out what maximizes expected utility relative to whatever body of information we are concerned with.[20]

---

which disagreement requires inconsistency. See e.g., Björnsson and Finlay (2010); Khoo and Knobe (forthcoming).

[19] This is the 'miners case', due to Parfit (2011, 1: p. 159). See discussion in Kolodny and MacFarlane (ms.); Björnsson and Finlay (2010).

[20] See, for example, Jackson (1991).

 © Pittard and Worsnip 2016

However, recent work elsewhere has suggested that *normative* uncertainty presents normative questions that may not be fully settled by first-order normative theories.[21] So far, the case of normative uncertainty has not been brought to bear on the debate surrounding 'ought' and contextualism. In this section, we argue that shifting to an example involving normative uncertainty rather than descriptive uncertainty allows for a clear and compelling argument in favour of metanormative contextualism over its rivals.

The example we will focus on concerns Greg, a policy maker who has the power to enact one of three proposed tax policies: policy A, policy B, and policy C. Policy A significantly increases the total well-being of the population as a whole, but does nothing to improve the lot of the worst off. Policy B significantly improves the lot of the worst off, but does not significantly affect the total well-being of the population. Policy C significantly improves the lot of the worst off, though not quite as much as policy B, and it also significantly increases the total well-being of the population as a whole, though not quite as much as policy A. Different theories of distributive justice will rank these policies differently. Utilitarianism holds that total well-being ought to be maximized and thus implies that Greg ought to choose policy A. Lexical prioritarianism holds that the well-being of the worst off trumps other considerations and thus implies that Greg ought to choose policy B. Finally, weighted prioritarianism holds that the worst off ought to receive some priority, but not the sort of absolute priority that completely trumps considerations of total well-being. Many versions of weighted prioritarianism will imply that Greg ought to choose policy C.

Suppose Greg is certain that either utilitarianism or lexical prioritarianism is correct, but that he is not especially confident in either of these theories. His credence for utilitarianism is, say, 0.51, and his credence for lexical prioritarianism is 0.49. Deliberating in his office, he says the following to his colleague: 'I know that the correct theory is either utilitarianism (in which case I ought to choose policy A) or lexical prioritarianism (in which case I ought to choose policy B). But since I have no idea which of these two theories is right, I ought to choose policy C. For on either theory, policy C is a close second'.

It is important not to misunderstand Greg's view here. His view is not that, since utilitarianism and lexical prioritarianism are both plausible, but somewhat extreme, theories, weighted prioritarianism

[21] See, e.g., Sepielli (2009), and the references therein.

 © Pittard and Worsnip 2016

is an even more plausible compromise between the two, and thus likely to be the true theory. Rather, he has credence o that weighted prioritarianism is actually the true theory of distributive justice: his whole credence space is divided between utilitarianism and lexical prioritarianism. (We can imagine that he has some particular reason for this: perhaps he thinks, for example, that weighted prioritarianism is very vague, and that no such vague theory could be the true theory.) Rather, his reason for following weighted prioritarianism is that he thinks it is a good compromise between the two theories he thinks may be correct, which is the appropriate way to act given his uncertainty. He chooses policy C because it is a close second on both of the metrics that he thinks could be correct.

Whether or not Greg expresses some mistaken views in his speech, the speech is entirely intelligible. Even if Greg is mistaken in thinking that either utilitarianism or lexical prioritarianism is correct, it is certainly intelligible for someone who is uncertain between these two views to think that the correct way to act in light of this uncertainty is to adopt some middle-ground action (one that happens to conform to some weighted prioritarian views). Of course, Greg might also be wrong about how to act in the face of uncertainty. Perhaps one should not adopt a course of action that gives some weight to both of the live possibilities, but should instead pick one policy (perhaps the one receiving the most credence) and perfectly conform to it. So Greg might be mistaken not only about his views on distributive justice, but also about the norms that describe how to act given uncertainty between these views. But mistaken or not, Greg's speech is understandable and seemingly coherent. It is not as though he has expressed some obviously inconsistent set of views.[22]

And yet Greg has said *both* that either he ought to choose policy A or he ought to choose policy B *and* that he ought to choose policy C (an action that is acknowledged to be incompatible with choosing policy A or choosing policy B). Thus, Greg's case provides a challenge for proponents of the competing semantic theories of 'ought'—

[22] In fact, it's plausible that whatever the reasonable credences regarding moral theories are, they do involve some amount of uncertainty; and that given that uncertainty, one ought to hedge one's bets to at least some extent. These claims are defended at length in the literature on moral uncertainty; see, e.g., Sepielli (2009). (Weatherson (2014) dissents here, but see §4 below.) But our argument does not actually depend on the claim that one ought sometimes to hedge one's bets in this way; all we need to claim is that Greg's utterances is understandable and not obviously unreasonable. Our argument starts from the apparent coherence of Greg's speech, not its truth.

 © Pittard and Worsnip 2016

namely, explaining how it is that Greg speaks in an at least apparently reasonable way despite making two 'ought' claims that, when 'ought' is assigned the same semantic value, appear to be inconsistent. We will show how contextualism accounts for the apparent coherence of Greg's speech before arguing that relativism, soft invariantism, and hard invariantism are all incapable of doing so.

3.1 Contextualism

Unsurprisingly, the idea behind the contextualist account of what is going on with Greg's speech is that his two uses of 'ought' have different semantic values. As a result, they do not end up contradicting each other. Crudely, one might say that Greg's claim that he either ought to choose policy A or ought to choose policy B is a claim about what he ideally ought to do given the actually true theory of distributive justice, whereas his claim that he ought to choose policy C is a claim about what he ought to do given his state of subjective uncertainty about the theory of distributive justice.

Again, there are different possible proposals about how to implement this semantically, as there were in the descriptive uncertainty case. Perhaps the most plausible proposal is that there is a shift in the relevant standards. The first part of Greg's speech could be interpreted as a claim about the correct choice in light of the norms of distributive justice, while the statement that he ought to choose policy C could be interpreted as a claim about the correct choice in light of the norms that govern how to act under normative uncertainty.

Some moderate contextualists who are happy to allow that the semantic value of 'ought' is relative to information might want to resist the idea that the relevant standards can vary across different contexts of utterance. They may feel that the latter kind of contextualism results in a theory that is too promiscuous with truth, allowing agents to speak truly when they make normative claims simply by subscribing to the relevant standards that vindicate those claims. Those of a realist metaethical persuasion, who think that there are *correct* and *incorrect* standards independently of anyone's holding those standards, may balk at this.

But this worry is misplaced, since standards-contextualism is on closer inspection fully compatible with this realist position. The standards-contextualist can allow that some normative speech 'aspires to objectivity' in that it is a claim about what one ought to do *given the actually true normative standards*, not about what one ought to do *given the speaker's standards*. In other words, the standards-

 © Pittard and Worsnip 2016

contextualist can deny that the standards parameter is always filled by the standards that the speaker subscribes to: when the speaker intends to make a more ambitious normative claim, the parameter can be filled by the actually true normative standards (if there be such).[23] Moreover, even if the standards-parameter is *never* filled simply by the subjective standards of the speaker, it may be filled by different kinds of standards on different occasions, compatibly with each set of standards being objectively correct. For instance, in Greg's case, it may be filled by the objectively true norms of distributive justice for one utterance, and by the objectively true norms of how to respond to normative uncertainty for the other utterance. So the choice between standards-invariantism and a view on which 'anything goes, always' is a false dichotomy.

However the contextualist semantics is to be precisely developed, it is clear that the contextualist needs to posit a change in context in the course of Greg's short speech that causes the semantic value of 'ought' to change. One might worry about this, given that Greg does not appear to have gained or lost any information in the course of his speech.[24] However, as we have just indicated, it is not obvious that the contextualist needs to understand the relevant parameter that shifts in value during Greg's speech as being information. More to the point, even if the relevant parameter is to be understood as information (specifically, *normative* information), recall the point we made in §1, namely that the contextualist is not committed to thinking that the information parameter is always filled by the information *available to* or *known by* the subject. When they intend to do so, it is possible for speakers to talk about what they ought to do relative to richer bodies of information, such as the full totality of the facts, even if these bodies of information are not (fully) available to them.

While the exact conversational mechanisms that govern shifts in contexts are hard to precisify, in Greg's case we do have a relatively clear marker of the context shift. Specifically, to introduce the

[23] This does not necessarily mean that the parameter is *always* filled by the actually true normative standards. On other occasions, the speaker might be making a less ambitious claim about what one ought to do given some set of standards that are assumed in the context.

[24] For various reasons, contextualists often hold that context-sensitive terms exhibit a certain 'stickiness', whereby it is difficult for either a single or multiple speakers to felicitously use the same context-sensitive term with differing semantic values in close succession, at least not without a clear signal that there has been a context shift in the relevant values of the contextual parameters that determine these semantic values. (See, e.g., DeRose 2004; Hannon 2015.)

 © Pittard and Worsnip 2016

'normative uncertainty "ought"', Greg says, 'since I have no idea which of these two theories is right…' This makes the perspective of normative uncertainty salient. Part of what governs shifts in context are the speaker's intentions to talk about different things, and here Greg signals an intention to talk about what he ought to do *given his uncertainty*.

Given this shift in context, what Greg says can be interpreted as consistent and cogent. It is important to see that the consistency contextualism wins us is not merely consistency in some impoverished sense, or in some specialized 'dynamic' or 'cross-context' sense. Utterances are inconsistent with one another in virtue of expressing inconsistent propositions. And when a term has different semantic values on different occasions of utterance, the propositions expressed are correspondingly different. So, if one says 'I ought to Φ' and also says 'I ought not to Φ', but 'ought' takes different semantic values on the two occasions, then the proposition one expresses with the second utterance is simply not inconsistent with the proposition one expresses with the first utterance. Any accusation of inconsistency is a simple mistake that is like the inverse of equivocation: it trades upon treating the term as having the same semantic value on the two occasions of utterance, when these semantic values are in fact different.

3.2 Relativism

One might be tempted to think that the relativist can simply mimic the contextualist's way of resolving the apparent inconsistency in Greg's speech. If the contextualist can say that the context of utterance changes during Greg's speech, why can't the relativist say that the context of *assessment* changes during Greg's speech?

It is not clear how to interpret this proposal. It is a determinate feature of any particular token utterance of a sentence that it occurs in some specific context. Individual token utterances in this sense *have* a context of utterance. When it comes to context-sensitive expressions, one can make a *mistake* by interpreting the utterance's meaning against the *wrong* context of utterance—that is, a different context of utterance than the one in which it is actually made. By contrast, individual token utterances do not *have* a context of assessment (MacFarlane 2014, pp. 61–2), and there are, on the relativist view, no semantic rules that determine a single, uniquely 'relevant' context of assessment from the context of utterance. Nor is one *required* to assess an utterance from any particular context of assessment, in the sense that one would be making a mistake by assessing it from some

 © Pittard and Worsnip 2016

different context of assessment. Rather, according to the relativist, a particular token utterance can be assessed for its truth-value at any number of different contexts of assessment. That is what makes the view distinctively relativist. Thus, while it is perfectly clear what it means to say that *the* context of utterance changes during Greg's speech, it is not at all clear what it means to say that *the* context of assessment changes during Greg's speech.

Perhaps what is meant is that the context of assessment that Greg himself occupies changes during his speech. When he says that he either ought to choose policy A or ought to choose policy B, he is projecting into the perspective of someone with full knowledge of the true norms of justice.[25] By contrast, when he says that he ought to choose policy C, he is speaking from his own context of normative uncertainty.

The first point to make in reply to this proposal is that, whatever it is supposed to achieve, there is still a crucial disanalogy between a contextualist appeal to a change in context of utterance, and a relativist appeal to a change in context of assessment. By appealing to a shift in the semantic value of 'ought', the contextualist explains how Greg does not actually express two inconsistent propositions. But whatever the appeal to a shift in context of assessment comes to, it will not deliver the result that Greg does not express inconsistent propositions. On the contrary, it is distinctive of the relativist view that utterances made by speakers occupying different contexts of assessment *do* express inconsistent propositions. In the case of Sarah and her professor, the relativist is at pains to stress that Sarah and her professor do express inconsistent propositions. Indeed, as noted in §2.3, relativists claim this as a data point in *favour* of their view, since it preserves the idea that there is a robust form of disagreement between the speakers. It is just that, from Sarah's context of assessment, her utterance is true and the professor's is false, whereas from the professor's context of assessment, his utterance is true and Sarah's is false.

So in a simple sense, the contextualist can deliver consistency to Greg's speech in a way that the relativist cannot. What the relativist could deliver, at most, would be that each utterance Greg makes is true at the context of assessment he occupies when he makes it. That does not obviously deliver the *coherence* of the speech as a whole. After

[25] Although, clearly he cannot do this fully, since if he could, he would know *which* of policies A and B to pick. But we will let this pass.

 © Pittard and Worsnip 2016

all, on the relativist account, from the context of assessment that Greg occupies when he says that he ought to choose policy A or he ought to choose policy B, the claim that he ought to choose policy C is false; whereas from the context of assessment that Greg occupies when he says that he ought to choose policy C, the claim that he ought to choose policy A or ought to choose policy B is false.

The relativist might initially be willing to bite that bullet. Perhaps, the relativist might say, Greg is just torn between two different perspectives here. Thus, the relativist might settle for the claim that each utterance Greg makes will seem true to him at the context of assessment he occupies when he makes it. Unfortunately, however, the relativist cannot even deliver this result. Consider Greg's speech again. What is striking about the speech is that it forms a complete piece of *reasoning*. Greg's reason for ultimately concluding, 'I ought to choose policy C' is that he is sure that *either* utilitarianism is correct (i.e., he ought to choose policy A) or lexical prioritarianism is correct (i.e., he ought to choose policy B), but uncertain as to which—and that policy C is a very close second on both metrics. His claim that he either ought to choose policy A or ought to choose policy B thus serves as a supporting *premise* in generating the claim that he ought to choose policy C. Yet, for the relativist, any context of assessment at which it is true that Greg ought to choose policy C must also be a context of assessment at which it is false that Greg either ought to choose policy A or ought to choose policy B. And if we interpret Greg as occupying a context of assessment relative to which it is false that he either ought to choose policy A or ought to choose policy B, then he loses his rationale for concluding that he ought to choose policy C. In other words, the relativist cannot identify a context of assessment from which Greg can reflectively endorse his claim that he ought to choose policy C. Consequently, his claim that he ought to choose policy C ceases to be rationally intelligible.

The contextualist has no analogous problem. On the contextualist's view, Greg uses the premise that he ought (on a more objective reading of 'ought', viz., what one ought to do given the true theory of justice) to choose policy A or he ought to choose B to generate the conclusion that he ought (on a more subjective reading of 'ought', viz., what one ought to do given one's normative uncertainty) to choose policy C. There is no problem with being able to simultaneously endorse both of these judgments from one perspective, because fundamentally they are consistent propositions that just happen to both use the same context-sensitive term, 'ought'. When you are

 © Pittard and Worsnip 2016

speaking at one context of utterance, you have no problem seeing the truth of claims made at different contexts of utterance, even if the sentences used to express those claims would be false uttered at your context of utterance. For you can see that the underlying propositions that those sentences express at their actual contexts of utterance are true. Nothing about your occupying your present context of utterance interferes with that judgment. By contrast, it's part of what it is to *occupy* a perspective or context of *assessment* in the relativist's sense that you assess other utterances by those standards too.

At this point, the relativist might protest that she has her own way of resolving these problems.[26] She might say that Greg can occupy a more detached perspective from which he acknowledges, compatibly, that relative to one context of assessment—CA1—he either ought to choose policy A or ought to choose policy B, but that relative to another context of assessment—CA2—he ought to choose policy C. And his speech and reasoning reflect his acknowledgment of both of these things.

The problem with this proposal is not that it is implausible but rather that, on closer inspection and development, it turns out to concede the truth of contextualism. Greg does not say anything explicit about different contexts of assessment in his speech. So to develop the current proposal to explain the apparent coherence of Greg's speech, we need to say that Greg's claim that he ought to choose policy A or he ought to choose policy B is a way of *talking about* what is true relative to CA1, and that his claim that he ought to choose policy C is a way of *talking about* what is true relative to CA2. But now we can see why this proposal is in fact a particular form of contextualism. For it effectively posits an unarticulated, contextually-determined parameter for the semantic value of 'ought' that could be made explicit by saying 'relative to context of assessment X…' Even though the content of this parameter concerns a context of assessment, the value of this parameter is supplied, on the present proposal, by the context of *utterance*, and as different contexts of utterance supply different implicit relativizations, the semantic value of 'Greg ought to Φ' will change.[27] That is a (particular kind of) contextualist view.

As we clarified at the start of this section, the relativist view as usually understood does not involve this contextualism: it is not

[26] Thanks to the editors of *Mind* for sharpening this objection.

[27] Arguably, Harman's (1975; 1996) version of 'moral relativism' is a disguised contextualism of this sort.

 © Pittard and Worsnip 2016

claimed that there is a uniquely relevant context of assessment for every token utterance, or that claims of the form 'you ought to Φ' *mean* 'relative to context of assessment CA, you ought to Φ'.[28] Rather, a single token utterance of the form 'you ought to Φ' can be assessed as true or false at any number of contexts of assessment. If the relativist switches to a view on which context of assessment features as a semantic parameter, the value of which is fixed by context of utterance, then she purchases a satisfactory handling of Greg's case only in virtue of conceding the truth of metanormative contextualism. And metanormative contextualism is what we have set out to establish with Greg's case.

As the above discussion helps to bring out, relativism and contextualism are not contradictory views, strictly speaking. Indeed, not only is relativism compatible with contextualism, but leading relativists such as Kolodny and MacFarlane acknowledge that the semantic value of 'ought' is context-sensitive in certain respects (2010, p. 131). For example, when someone says, after looking at a certain chess position, that Black ought to move her queen (in order to avoid a needless queen sacrifice), she does not contradict the father who, when confronting this same position in a game against his young son, says to himself, 'I ought to leave my queen where she is, in order to give my son an opportunity to win the match'. Kolodny and MacFarlane would acknowledge that these two 'ought' claims are not inconsistent, since the contextually relevant 'deontic selection function' (to use their terminology) differs across the two utterances. In the first utterance, the relevant selection function evaluates options in light of the goal of placing white in checkmate, while in the second utterance the relevant selection function evaluates options according to some other criteria (e.g., moral criteria, or the criteria for being a good parent).

So Kolodny and MacFarlane do not oppose contextualism per se. What they oppose is a contextualist resolution to the puzzle presented by cases like Sarah the herbalist. One who seeks to explain the herbalist case in contextualist terms must endorse a version of contextualism that is more thoroughgoing than the restricted version admitted by Kolodny and MacFarlane. According to this more thoroughgoing contextualism, the semantic value of 'ought' is not fixed once it is

[28] Indeed, as we have already noted, one of the central arguments for relativism—that those speaking from different contexts nevertheless disagree with one another in a robust sense involving inconsistency—crucially turns on the *rejection* of such a view. If two speakers are each only talking about what is true at their own contexts of assessment, and their contexts of assessment differ, there is no longer any inconsistency in what they say.

 © Pittard and Worsnip 2016

determined that the speakers are concerned with the 'ought' of morality (or of chess, or prudence, or what have you). Even when restricting ourselves to the 'ought' of morality, for example, determining the semantic value will depend on further parameters (e.g., information) that vary with context. The contextualism required to explain Greg's case goes much further than merely acknowledging the difference between the 'ought' of morality and the 'ought' of chess: the use of 'ought' to convey the normative implications of various states of uncertainty requires that the semantic value of 'ought' vary in a more systematic way that is along the lines of the contextualism posited to explain the case of Sarah the herbalist.[29]

Kolodny and MacFarlane reject this thoroughgoing contextualism and argue that relativism can explain why Sarah and the professor both speak reasonably in the herbalist case. But as we've argued here, relativism cannot explain the reasonableness and apparent coherence of Greg's speech. This does not by itself show that relativism is false, but only shows that the relativist will either need to invoke a more thoroughgoing version of contextualism (thus conceding our argument) or appeal to one of the invariantist strategies. Moreover, to the extent that we should prefer a unified solution to the case of Greg and that of Sarah—cases that, notwithstanding the difference between normative and descriptive uncertainty, have some striking similarities—the case of Greg puts pressure on the relativist proposal about Sarah's case.

We now argue that the invariantist strategies also fail to explain the apparent coherence of Greg's speech, leaving metanormative contextualism as the only viable option.

3.3 Hard invariantism

Recall the two-pronged strategy that we sketched for the hard invariantist in §2.2 for dealing with cases of descriptive uncertainty: either say that the 'ought' under uncertainty takes wide-scope, or say that the 'ought' under certainty is simply false and a result of mistaken

[29] Admittedly, Kolodny and MacFarlane themselves only commit to a relativist treatment of the informational component of the semantics. So if we are determined to understand descriptive uncertainty as 'informational' and normative uncertainty as 'non-informational' (though this is a questionable bifurcation), it may look like they are not committed to a relativist treatment of normative uncertainty. Since they are silent on normative uncertainty, it is not entirely clear what their treatment of it would be. Perhaps they would concede a fully contextualist treatment of normative uncertainty. But the point remains that this would amount to conceding a much more thoroughgoing contextualism than they have seen the need to admit on other grounds.

 © Pittard and Worsnip 2016

projection. Does either of these options help with examples like that of Greg?

As we suggested earlier, the wide-scoping option is at its most plausible when the uncertainty on the part of the agent is not in fact justified. According to the wide-scoping option, the most charitable way to interpret Greg's beliefs involves ascribing to him (i) the belief that he ought to choose policy A or he ought to choose policy B, and (ii) the (wide-scope) belief that he ought, if he is uncertain about which of policy A or policy B he ought to choose, to choose policy C. If Greg believes that his own uncertainty about whether he ought to choose policy A or ought to choose policy B is unjustified (or, indeed, if he just fails to believe that his uncertainty *is* justified), then it is clear that he can coherently believe (i) and (ii) together. But just as in the case of Sarah, it is easy to imagine the case so that Greg believes that his normative uncertainty is justified, indeed even required (that is, that he ought to be uncertain). Adding this stipulation does not in any way undercut the impression that Greg's speech is reasonable and coherent, which suggests that the apparent coherence of Greg's speech does not derive from an implicit assumption that he believes that his uncertainty is unjustified.

Note that simply denying that normative uncertainty is ever justified, as some philosophers are wont to do, is not sufficient to defend the wide-scoping option, at least so long as Greg himself thinks that his normative uncertainty is justified.[30] Nor is it likely that when typical hearers perceive Greg's speech as coherent, they are tacitly assuming that his normative uncertainty is unjustified. If it is a fact that normative uncertainty is never justified, it is a surprising one and not one that is in the background when typical hearers interpret Greg's speech as coherent. A viable *semantic* thesis should help to explain why Greg and others *take* his speech to be a reasonable one.

Another suggestion aimed at rescuing the wide-scope variant of hard invariantism would be that Greg's beliefs are unproblematically coherent because *deontic dilemmas* are possible. For our purposes, an agent counts as being in a deontic dilemma just in case it is true from a single context of assessment, and without any variation in the semantic value of 'ought', that the states of affairs he ought to bring about are collectively incompatible. If we stipulate that Greg believes

[30] For arguments that appear to support the surprising conclusion that normative uncertainty is never justified, see Kiesewetter (2016) and Titelbaum (2015), though the latter restricts his thesis to beliefs about *rationality*.

 © Pittard and Worsnip 2016

that his uncertainty is required, and if we accept what the hard invariantist who takes the wide-scoping option has to say about Greg's beliefs, then the only way for Greg's package of beliefs to come out true would be if Greg is in a dilemma of this sort.

This is easy to see if deontic detachment (introduced in §2.2 above) is correct. In this case, the wide-scope claim, together with the claim that Greg ought to be uncertain about whether he ought to choose policy A or ought to choose policy B, would entail that Greg ought to choose policy C, an action that is incompatible both with his choosing policy A and with his choosing policy B. One might think that if we deny deontic detachment, the deontic dilemma disappears. But in fact this is not so.[31] Acting so as to ensure the truth of the conditional that [if Greg is uncertain about which of policy A or B he ought to choose, then he chooses policy C] requires ensuring either that Greg chooses policy C (which is incompatible with choosing policy A and with choosing policy B) *or* that he is not uncertain about which of policy A or policy B he ought to choose (which is incompatible with the claim that he ought to be uncertain). Since on the wide-scoping option for hard invariantism Greg's package of beliefs would imply that he is in a deontic dilemma, a natural move in the effort to explain the apparent coherence of Greg's speech would be to appeal to the coherence of the claim that Greg's situation constitutes a deontic dilemma.[32]

Is it plausible to maintain that the apparent coherence of Greg's speech is explained by our recognizing at some level the plausibility of deontic dilemmas? We don't think so. For starters, whether Greg's speech appears coherent does not seem to depend on whether one finds deontic dilemmas plausible. (We, at least, find Greg's speech to be eminently reasonable even though we are not inclined to endorse deontic dilemmas.) Indeed, again, it does not seem to be a presupposition of his speech that *he himself* (even implicitly) takes himself to be in a deontic dilemma, which suggests that this is not what explains his speech's apparent coherence. Moreover, if Greg's hearers implicitly understood him to be describing a dilemmatic situation, then they would likely be left wondering what he will in fact do. But in fact, after

[31] Thanks to an anonymous referee here. Interestingly, though Broome denies deontic detachment (as noted earlier), he seems hostile to the possibility at least of *rational* dilemmas (Broome 2007, p. 364).

[32] A hard invariantist might alternatively make the proposal that Greg is in a deontic dilemma quite independently of the wide-scoping strategy. Our objections here apply to either version of the proposal.

 © Pittard and Worsnip 2016

hearing Greg's speech, with its superficially incompatible 'ought' claims, it is clear that Greg has resolved to choose policy C. The speech exhibits reasoning that resolves the question of how to act rather than reasoning that shows why Greg is torn between incompatible requirements and condemned to moral failure. Finally, and relatedly, the two judgments—that he either ought to choose policy A or ought to choose policy B, and that he ought to choose policy C—do not appear to be putting any kind of pressure on each other, as conflicting 'ought'-judgments do in cases that are typically taken as candidates to be deontic dilemmas.[33] (Indeed, as noted in the discussion of relativism above, the former judgment actually serves as the *grounds* for the latter.) This strongly suggests that the two 'ought's are in fact taking different semantic values, rather than a single value that is giving rise to a deontic dilemma.

The discussion thus far gives us reason for thinking that the first hard invariantist option, appealing to a wide-scope normative belief, will not adequately explain the apparent coherence of Greg's speech. That leaves us with the second prong of the hard invariantist strategy, namely that the 'ought' under certainty is just a mistaken projection. Obviously, the notion of mistakenly projecting makes more sense when we have one speaker who projects his own epistemic position onto that of a different subject, as the professor does. But Greg cannot be doing this to himself. So the best the hard invariantist can do here is to bite the bullet and simply maintain that Greg's claim that either he ought to choose policy A or he ought to choose policy B is just false, and that it is guaranteed to be so by his claim that he ought to choose policy C.

But again, this suggests that Greg's speech should appear incoherent, for the same reason we gave in response to the relativist. If Greg's claim that he ought to pursue policy C entails that he speaks falsely in saying that he ought to choose policy A or he ought to choose policy B, then his reasoning for the former claim is self-undermining, since the latter claim is a premise in his argument for it. Consequently, this version of the hard invariantist view makes him blatantly incoherent.

[33] Of course, the advocate of deontic dilemmas holds that in such cases, ultimately the two conflicting 'ought' claims are logically consistent. But they still put at least some *prima facie* epistemic pressure on each other. For example, in Sartre's (1946/2007) famous example, the judgment that one should look after one's mother puts at least *some* epistemic pressure on the judgment that one should fight for the free French, and vice versa. Each provides at least some epistemic reason to doubt the other, even if we ultimately conclude that both judgments are true.

 © Pittard and Worsnip 2016

Could the hard invariantist maintain instead that it is Greg's claim that he ought to choose policy C that is false?[34] So-called 'objectivists' hold that what one ought to do (in any legitimate, genuine sense) is simply not sensitive to one's epistemic position (or one's uncertainty),[35] and so will maintain that there is no legitimate sense in which Greg ought to choose policy C. But on closer inspection, this first-order normative view has no bearing on the debate between hard invariantism and contextualism, which are *semantic* theories. The fact that Greg speaks as he does, and seems coherent in doing so, still suggests that he is using 'ought' in a way that presupposes that differences in epistemic position can make a difference to what one ought (in at least some sense) to do. If objectivism is true, then this presupposition is false, and so claims that involve such uses of 'ought' are false (or, at least, suffer from presupposition failure). But that does not mean that they are not *using* 'ought' in the way we have described. The objectivist view can yield an error theory about non-objective uses of 'ought', rather than denying that there *are* such uses.[36]

Of course, an objectivist *could* instead say that all usages of 'ought' do in fact express the objective, uncertainty-insensitive 'ought'. But this theory simply gives no explanation for the apparent coherence of Greg's speech (and, more broadly, the wide range of linguistic data suggesting that people do in fact use 'ought' in a (putatively) uncertainty-sensitive way). So we think the error-theoretic version of objectivism, on which it is married with contextualism at the semantic level, is more promising. Again, our claim was not that Greg's speech is true but only that it is (at least apparently) coherent. The error-theoretic objectivist, by accepting contextualism, can agree with us that Greg's speech is coherent, and that his two claims do not contradict each other. She just goes on to say that one of them, given the

[34] Thanks again to an anonymous referee for this suggestion.

[35] See e.g., Graham (2010).

[36] Compare more familiar error theories about the moral 'ought' of the sort defended by Mackie (1977) and Joyce (2001). Joyce is explicit (and here he is drawing out what is plausibly latent in Mackie also) that he thinks there are true claims that use 'ought', so long as such claims only use the *rational* 'ought' as opposed to the moral ought. But his strategy is not to claim that every usage of 'ought' actually expresses the rational 'ought'. On the contrary: what makes him an error theorist is precisely the fact that he thinks that much ordinary usage of 'ought' *does* clearly employ the moral 'ought', presupposing that there are objective, categorical moral standards of the sort that Joyce thinks there cannot be. Consequently, on Joyce's view, much ordinary normative talk turns out to be false. Similarly, the objectivist can recognize that there are subjective or information-sensitive uses of 'ought' but be an error theorist about such uses, holding that they result in claims that are false (or suffer from presupposition failure).

 © Pittard and Worsnip 2016

normative truth of objectivism, is false. While we ourselves do not accept objectivism, our argument does not turn on its rejection.

3.4 Soft invariantism

Finally, we turn to soft invariantism. Recall from §2.1 that attempts to appeal to ambiguity between different *normative* 'ought's—for example, between 'objective' and 'subjective' 'ought's—are subject to a severe regress problem. Since one can be uncertain about what one ought to do even in the more 'subjective' sense posited, the soft invariantist is forced to posit a third 'ought', which in turn can be attended by uncertainty, forcing a fourth 'ought', and so on. When the semantic values of 'ought' begin to multiply in this way, it is clear that these various semantic values possess an underlying unity that calls for a contextualist semantics rather than a proliferation of multiple senses.

This problem applies equally when it is normative uncertainty that is at issue. One can be uncertain, not just about what the true theory of normative ethics is, but about what the true theory of how one ought to act under normative uncertainty is, and about what the true theory of how one ought to act under uncertainty about the true theory of how one ought to act under normative uncertainty is, and so on.[37]

It's worth noting that this problem also arises for another proposal that may occur to a soft invariantist in the case of normative uncertainty—namely, to try to distinguish between a 'moral sense' of 'ought' and a 'rational sense' of 'ought'.[38] Suppose one says that morally, Greg ought to choose whichever policy is actually recommended by the true theory of distributive justice, but rationally, he ought to choose policy C. This is no better than the binary objective/subjective distinction. Again, Greg can be uncertain about how he ought to act under normative uncertainty, and that forces this particular kind of soft invariantist to admit a proliferation of senses of the rational 'ought', just as it forced our original soft invariantist to admit a proliferation of senses of the subjective 'ought'.[39]

[37] See also Sepielli (2014).

[38] Of course, the contextualist can recognize a difference between moral uses of 'ought' and rational uses of 'ought'. What follows is not an argument against distinguishing the moral and rational 'ought' per se, but rather against trying to use this distinction to resist contextualism.

[39] A soft invariantist might try replying to this by claiming that the 'rational "ought"' is not actually a normative 'ought'. In this way, like the soft invariantist who claims that only the evaluative 'ought' is context-sensitive, she avoids admitting contextualism about the *normative* 'ought'. Inspiration for this proposal might be drawn from Kolodny (2005), who argues that

 © Pittard and Worsnip 2016

In the case of descriptive uncertainty, however, recall that there was a different soft invariantist strategy that avoided these problems. This strategy was to maintain that the sense in which Sarah ought to have given herb B was some non-normative sense (unlike the sense in which she ought to have given herb C). So when the professor says that Sarah ought to have given herb B, this 'ought' claim does not express a normative view about what Sarah had (or has) most reason to do, but rather (for example) an *evaluative* view about the optimific action. So even if the soft-invariantist is forced to concede that 'ought' can be context-sensitive, she can seek to preserve an invariantist semantics for the *normative* 'ought' by insisting that any context-sensitivity is limited to the non-normative senses of 'ought'.

While this sort of soft invariantism is, we think, relatively plausible in the case of descriptive uncertainty, it is much less plausible in the case of normative uncertainty. Normative uncertainty is, after all, uncertainty about the truth of some *normative* theory. In the case of Greg, for example, utilitarianism and lexical prioritarianism are opposing normative theories of how one ought to distribute goods, and so the 'ought' claims they issue must be *normative* ought claims. (By contrast, remember that Greg *denies* that weighted prioritarianism is the true theory of distributive justice.) So when Greg sees that utilitarianism implies that he ought to choose policy A and that lexical prioritarianism implies that he ought to choose policy B, what he is seeing (assuming he is not confused) are normative implications of these two theories. Thus, a reasonable interpretation of Greg's speech requires that we understand his claim that either he ought to choose policy A or he ought to choose B as using 'ought' in a normative sense.

The soft invariantist might reply that, contrary to appearances, utilitarianism and lexical prioritarianism are evaluative theories of the optimific state of world, and not of what one ought to do in the

---

the requirements of rationality are not normative. However, the 'requirements of rationality' that Kolodny denies to be normative are pure *coherence* requirements. Central to Kolodny's argument is the idea that coherence never enters into deliberation: one does not think 'I ought to Φ, because Φ-ing will make me rationally coherent'; rather, one thinks, 'I ought to Φ, because Φ-ing would be good, or right, or get me what I want'—and coherence comes about at most as a by-product of this kind of good deliberation. Hence, any sense in which one ought to be coherent *as such* is (arguably) a non-normative sense. But it is clear that the same argument does not apply merely to 'ought's under normative uncertainty. On the contrary, for Greg, the 'ought' under normative uncertainty plays a *crucial* role in his deliberation. So if what is meant by 'rational "ought"' is '"ought" under (normative) uncertainty' rather than '"ought" of satisfaction of coherence requirements', there seems to be little scope to deny the normativity of this 'ought'.

 © Pittard and Worsnip 2016

normative sense. We think this is *ad hoc*. Utilitarianism and lexical prioritarianism are commonly stated as normative theories about what actions one ought to perform. And even if the soft invariantist claims that these theories would be better understood as evaluative theories, it is still not clear why *Greg* need be thinking of them as evaluative when he is uncertain between them.

Moreover, a cousin of the regress problem that doomed other soft invariantist accounts also looms here. On the particular elaboration of the soft invariantist view that we are currently considering, it turns out that the theories of so-called 'normative ethics' are not actually normative but evaluative. The truly normative theory is a theory about how to act under uncertainty about which of these theories is correct. Call this a theory of 'first-order' uncertainty. But one can, in turn, be uncertain about which theory of first-order uncertainty is correct.[40] In such a case, we could talk about how one ought to act given the true theory of how to act under first-order uncertainty. But we could also talk about how one ought to act given one's uncertainty about the true theory of how to act under (first-order) uncertainty. The latter would constitute a theory of second-order uncertainty.

Just as the soft invariantist blocked contextualism by saying that theories of normative ethics are not really normative, it now seems that she must block contextualism by saying that theories of first-order uncertainty are not really normative. After all, one can be uncertain about them, and there are theories about how one ought to act given this uncertainty. The truly normative theories, then, are the theories of how to act under 'second-order' uncertainty. But of course, one can also be uncertain about the correct theory of second-order uncertainty. And so the regress keeps on going. As long as it is possible to be uncertain about the correct theory of how to act under some level of *n*-th order uncertainty, we can cook up a case to show that the soft invariantist must claim that theories of *n*-th order uncertainty are not normative. There are, then, no stably normative theories for this kind of soft invariantist.

[40] For example, in addition to his uncertainty between utilitarianism and lexical prioritarianism, Greg might be uncertain between two opposing theories of how one ought to choose when uncertain between competing theories of distributive justice: a 'winner-takes-all' view on which one ought to choose the policy that accords with the theory for which one has the highest credence; and a 'compromise' view on which (when possible) one ought to choose a policy like C that is ranked highly according to each theory (even if it is not the policy required by either theory).

 © Pittard and Worsnip 2016

Perhaps an interesting defence of this radical idea could be made. However, we think that it is probably not what soft invariantists had in mind from the start. Ultimately, metanormative contextualism probably saves more of what typical soft invariantists wanted than this radical view does. After all, metanormative contextualism and soft invariantism share the contention that 'ought' takes different semantic values on different occasions of utterance. But contextualism can unify these different semantic values under a single semantics rather than positing ambiguity.

4. The relevance of *normative* uncertainty

Observant readers will have noticed that there are really two important differences between the cases of Sarah and Greg. First, there is the fact that Greg's uncertainty is normative, whereas Sarah's is descriptive. But second, there is the fact that Greg's case involves a *single speaker* making the two contrasting 'ought' claims, whereas Sarah's involves two *different* speakers. One might naturally be suspicious that it is really the single-speaker feature of Greg's case that is making all the difference in securing the case for metanormative contextualism. And clearly it does play some role. However, we also think that it is crucial that the case involves normative uncertainty. To show this, let us consider a single-speaker version of Sarah's case, and briefly illustrate why it does not tell as decisively in favour of contextualism.

Suppose that as Sarah is preparing a herb C concoction to give to Sam, she says the following: 'What I *really* ought to do is give you the herb that has no side effects. But I can't remember whether that's herb A or B, and giving you the wrong one would kill you. Given this uncertainty, I ought to give you herb C instead. It will save your life and leave you with only mild hypersensitivity to the sun'. Again, Sarah's initial 'ought' claim (that what she really ought to do is give the herb with no side effects) superficially appears to be inconsistent with the 'ought' claim made two sentences later (that she ought to give herb C instead).

The crucial difference between this case and Greg's, however, is that Sarah can and does reason to the conclusion that she ought to give herb C without relying on the premise that she ought to give herb A or B. She can reason to this conclusion purely from the descriptive facts that one of herb A and herb B will kill Sam, and that she can't remember which. Greg, by contrast, reasons to the conclusion that he

 © Pittard and Worsnip 2016

ought to perform policy C from the *normative* premise that either he ought to choose policy A or he ought to policy B, but doesn't know which.

As we saw earlier, this difference is important in arguing against both the relativist and hard invariantist stances. Both relativists and hard invariantists portray Greg's claim that he ought to choose policy A or he ought to choose policy B as inconsistent with his claim that he ought to choose policy C. But if they are inconsistent, then the former cannot serve as Greg's rationale for the latter, since he cannot endorse both simultaneously, and so his judgment that he ought to perform policy C is not intelligible. By contrast, even if we cut out any claim on Sarah's part that she ought to give the herb with no side effects, her claim that she ought to give herb C is fully intelligible. So it is easier to say that the two claims are actually inconsistent while preserving the intelligibility of the second claim, since in Sarah's case the second claim does not depend for its justification on the first claim.

The significance of introducing normative uncertainty is further underscored when we consider the error theory offered by relativists Kolodny and MacFarlane (2010) as an explanation for why Sarah might mistakenly affirm that she ideally ought to give the herb without side effects. Even if their error theory offers a plausible explanation in the example of Sarah, the error theory fails to apply to Greg's case. According to Kolodny and MacFarlane, Sarah can appropriately make the following claims, each of which is true at her context of assessment:

(i) If herb A is the optimal herb, then I ought to give herb A.

(ii) If herb B is the optimal herb, then I ought to give herb B.

(iii) Either herb A is the optimal herb or herb B is the optimal herb.

It would seem that these three claims logically entail a fourth claim:

(iv) Either I ought to give herb A *or* I ought to give herb B.

On the supposition that (iv) is logically entailed by (i)-(iii), then we would have to affirm that (iv) is also true at Sarah's context of assessment, which is of course what Kolodny and MacFarlane aim to deny. But Kolodny and MacFarlane maintain that (i)-(iii) do *not* logically entail (iv), since according to their relativist semantics *modus ponens* is not logically valid for conditionals that (like (i) and (ii) above) have

 © Pittard and Worsnip 2016

information-sensitive terms in the consequent. Because *modus ponens* is not valid, Sarah is unable to validly infer (iv) from (i)-(iii), and thus cannot validly arrive at a conclusion that is inconsistent with the claim that she ought to give herb C. But the reasons that account for the failure of *modus ponens* are quite subtle, which explains why Sarah would make the mistake of voicing (iv) in situations when she is entitled to assert (i)-(iii). We thus have an explanation for why we are inclined to endorse as appropriate Sarah's assertion of (iv), despite the falsity of (iv) at her context of assessment.

But when we turn to the case of Greg, the relativist cannot appeal to *modus ponens* failure in order to explain why Greg makes inconsistent claims. To see this, let's first see how Kolodny and MacFarlane's strategy would translate to Greg's case. Kolodny and MacFarlane affirm that Greg can appropriately make each of the following claims, each of which may be true at his context of assessment:

(i*) If utilitarianism is true, I ought to choose policy A.

(ii*) If lexical prioritarianism is true, I ought to choose policy B.

(iii*) Either utilitarianism is true or lexical prioritarianism is true.

These three claims would seem to entail the following:

(iv*) Either I ought to choose policy A or I ought to choose policy B.

Because (iv*) seems to logically follow from (i*)-(iii*), we have an explanation for why Greg is inclined to affirm it (and why we are inclined to judge that this affirmation is appropriate). But because Kolodny and MacFarlane hold that *modus ponens* is invalid in this setting, they would deny that (iv*) follows from (i*)-(iii*). This provides an explanation for why Greg (supposedly) cannot properly affirm (iv*), a claim that is inconsistent with his claim that he ought to choose policy C.

Unfortunately for the relativist, however, the claim that Greg cannot affirm (iv*) does not save Greg from inconsistency entirely. This is because (iii*), itself being a normative claim, is *already* inconsistent with Greg's claim that he ought to choose policy C. In Sarah's case, (i)-(iii) contain only *conditional* normative claims. *Modus ponens* is needed to extract an unconditional normative claim, (iv), that can conflict with Sarah's claim that she ought to give herb C. And since

 © Pittard and Worsnip 2016

this particular instance of *modus ponens* is supposedly invalid, no conflict can be generated. In Greg's case, however, (iii*) is an unconditional normative claim (unlike the merely descriptive (iii)). Because (iii*) is already normative in character, we do not need to use *modus ponens* in order to extract a claim that is inconsistent with Greg's verdict that he ought to choose policy C. (iii*) itself is inconsistent with this verdict because (iii*) on its own entails (without recourse to *modus ponens*) that either one ought to maximize total well-being or one ought to do the action that is best for the least well off.

The soft invariantist strategy is also much better off in the case of descriptive uncertainty than that of normative uncertainty, even when both are taken in their single-speaker form. The soft invariantist can quite plausibly say that Sarah's claim that she ought to give herb A or herb B, like the professor's claim in the two-speaker case, is really an evaluative claim about what would be optimific. But, as we have already argued, this is far less plausible in Greg's case, where his claim that he ought to enact policy A or policy B is a result of his certainty that one of two *normative* theories of distributive justice is correct. So, again, the case of normative uncertainty is much stronger for the contextualist than the case of descriptive uncertainty.

We have been suggesting that the phenomenon of normative uncertainty bears upon the debate between contextualist accounts of the normative 'ought' and their rivals. However, it may also be that the viability of metanormative contextualism is significant for debates over the normative significance of normative uncertainty. Without a contextualist semantics for the normative 'ought', we have argued, moral reasoning of the sort voiced by Greg would be positively incoherent. Yet those who affirm that moral uncertainty matters morally appear to endorse the sort of reasoning that Greg espouses. After all, they take the view that it can happen that what you ought to do is *not* just whatever the correct moral theory says, but rather some action that takes account of your uncertainty between plausible moral theories. This contention can appear to amount to saying that you ought not to do what you actually ought to do, which sounds incoherent.

Weatherson (2014, p. 146) has recently noted the apparent incoherence of this sort of response to moral uncertainty, and partly on this basis concludes that moral uncertainty has no bearing on what one ought morally to do. But if metanormative contextualism is correct, then this type of reasoning is not incoherent after all, at least not in any straightforward way. As we have been suggesting, such reasoning, at least in Greg's case, does not *appear* incoherent, and none of the

 © Pittard and Worsnip 2016

rivals to contextualism can explain even this *apparent* coherence. So we think it is better to conclude (with contextualists) that reasoning like it is not *in fact* incoherent, than to conclude (with Weatherson) that reasoning like Greg's must be incoherent and to deny the normative significance of normative uncertainty on these grounds. This of course does not establish that moral uncertainty has normative significance, but it does block one sort of incoherence argument that has been used to suggest that it does not.

5. Conclusion

Metanormative contextualism is the view that the semantic value of a given 'ought' claim depends on the value of one or more parameters that are fixed by the context of utterance. So, if there is even a single case that shows that 'ought' can have different semantic values in a way that cannot be explained by a soft invariantist appeal to 'multiple senses', then metanormative contextualism follows. We have argued that a (single-speaker) case featuring normative (rather than descriptive) uncertainty is such a case. This type of case establishes the truth of metanormative contextualism even if in other cases explaining the data requires appealing to relativism or one of the invariantist strategies assessed here.

That said, there is reason to prefer a unified explanation of these many cases. So even if the relativist or invariantist can do just as well as the contextualist at explaining some particular cases, the clear superiority of contextualism in the kind of case developed here gives us a reason for preferring the contextualist explanation in cases where contextualism does not have such a clear explanatory advantage. So, while our argument does not provide a knock-down argument that relativism is false—for contextualism and relativism could both be true—it puts considerable pressure on the attempt to motivate relativism with recourse even to the 'stalemate' cases such as that of Sarah. For relativism is unable to explain the nearby and structurally similar case that we have offered.

The invariantist is in one way better off than the relativist and in one way worse off. She is worse off because if contextualism is true, it simply follows that invariantism is false, given how we defined 'invariantism'. But many of the phenomena that the invariantist used to explain away the existing contextualist data may yet be genuine ones. And in this respect, she is perhaps better off than the relativist, for

 © Pittard and Worsnip 2016

there are considerable independent reasons to think both that, as the soft invariantist argues, 'ought' can have 'multiple senses' (normative, evaluative, and so on), and that, as the hard invariantist argues, it can sometimes take wide scope over conditionals (and express true claims in so doing). Nevertheless, we have argued that there is a particular kind of datum that only contextualism can account for, and thus that contextualism is true.

There is a huge amount of work to be done in precisifying metanormative contextualism, and we have done virtually none of it here. Rather, we have tried to argue that some form of it must be correct, however the details are to be spelled out, and however the challenges are to be overcome. Since metanormative contextualism has the capacity to shed light on many debates and problems in normative and metanormative philosophy, that should be welcome news.[41]

References

Björnsson, Gunnar, and Stephen Finlay 2010. 'Metaethical Contextualism Defended'. *Ethics* 121, pp. 7–36.

Brogaard, Berit 2008. 'Moral Contextualism and Moral Relativism'. *The Philosophical Quarterly* 58, pp. 385–409.

Broome, John 1999. 'Normative Requirements'. *Ratio* 12, pp. 398–419.

——2004. 'Reasons'. In *Reason and Value: Themes From the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, pp. 28–55. Oxford: Oxford University Press.

——2007. 'Wide or Narrow Scope?' *Mind* 116, pp. 359–70.

——2013. *Rationality through Reasoning*. Chichester: Wiley Blackwell.

Brunero, John 2010. 'The Scope of Rational Requirements'. *The Philosophical Quarterly* 60, pp. 28–49.

——2012. 'Instrumental Rationality, Symmetry and Scope'. *Philosophical Studies* 157, pp. 125–40.

Cariani, Fabrizio 2013. '"Ought" and Resolution Semantics'. *Noûs* 47, pp. 534–58.

Cariani, Fabrizio, Magdalena Kaufmann, and Stefan Kaufmann 2013. 'Deliberative Modality under Epistemic Uncertainty'. *Linguistics and Philosophy* 36, pp. 225–59.

[41] The authors contributed equally to this manuscript. For helpful discussion and/or comments we are grateful to Gunnar Björnsson, John Broome, Keith DeRose, Dan Greco, Josh Knobe, Ralph Wedgwood, the editors at *Mind*, and several anonymous referees.

 © Pittard and Worsnip 2016

Dancy, Jonathan 1977. 'The Logical Conscience'. *Analysis* 37, pp. 81–4.

——2000. *Practical Reality*. New York: Oxford University Press.

DeRose, Keith 2004. 'Single Scoreboard Semantics'. *Philosophical Studies* 119, pp. 1–21.

Dowell, Janice 2012. 'Contextualist Solutions to Three Puzzles about Practical Conditionals'. *Oxford Studies in Metaethics* 7, pp. 271–303.

——2013. 'Flexible Contextualism about Deontic Modals: A Puzzle about Information-Sensitivity'. *Inquiry* 56, pp. 149–78.

Dreier, James 1990. 'Internalism and Speaker Relativism'. *Ethics* 101, pp. 6–26.

Feldman, Richard 1988. 'Subjective and Objective Justification in Ethics and Epistemology'. *The Monist* 71, pp. 405–19.

Finlay, Stephen 2014. *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

van Fraassen, Bas C. 1972. 'The Logic of Conditional Obligation'. *Journal of Philosophical Logic* 1, pp. 417–38.

Gibbard, Allan 2005. 'Truth and Correct Belief'. *Philosophical Issues* 15, pp. 338–50.

Graham, Peter A. 2010. 'In Defence of Objectivism about Moral Obligation'. *Ethics* 121 (1): pp. 88–115.

Greenspan, Patricia S. 1975. 'Conditional Oughts and Hypothetical Imperatives'. *The Journal of Philosophy* 72, pp. 259–76.

Hannon, Michael 2015. 'Stabilizing Knowledge'. *Pacific Philosophical Quarterly* 96 (1): pp. 116–39.

Hansson, Bengt 1969. 'An Analysis of Some Deontic Logics'. *Noûs* 3, pp. 373–98.

Hare, Richard M. 1971. 'Wanting: Some Pitfalls', in his *Practical Inferences*, pp. 44–58. London: Macmillan.

Harman, Gilbert 1975. 'Moral Relativism Defended'. *The Philosophical Review* 84, pp. 3–22.

——1996. 'Moral Relativism'. In *Moral Relativism and Moral Objectivity*, by Gilbert Harman and Judith Jarvis Thomson. Cambridge, Mass.: Blackwell.

Henning, Tim 2014. 'Normative Reasons Contextualism'. *Philosophy and Phenomenological Research* 88, pp. 593–624.

Jackson, Frank 1991. 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection'. *Ethics* 101, pp. 461–82.

Joyce, Richard 2001. *The Myth of Morality*. Cambridge, UK: Cambridge University Press.

 © Pittard and Worsnip 2016

Kagan, Shelly ms. 'The Paradox of Methods'.
Khoo, Justin and Joshua Knobe, forthcoming. 'Moral Disagreement and Moral Semantics', *Noûs*, doi: 10.1111/nous.12151.
Kiesewetter, Benjamin 2016. 'You Ought to Φ Only If You May Believe That You Ought to Φ'. *The Philosophical Quarterly* 66, pp. 760–82.
Kolodny, Niko 2005. 'Why Be Rational?' *Mind* 114, pp. 509–63.
Kolodny, Niko, and John MacFarlane ms. 'Ought: Between Objective and Subjective'. http://phil.unlv.edu/ought-unlv.pdf.
—— 2010. 'Ifs and Oughts'. *The Journal of Philosophy* 107, pp. 115–43.
Kratzer, Angelika 1981. 'The Notional Category of Modality'. In *Words, Worlds, and Contexts*, edited by H. J. Eikmeyer and H. Rieser, pp. 38–74. Berlin: de Gruyter.
Lewis, David 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.
Lord, Errol 2011. 'Violating Requirements, Exiting from Requirements, and the Scope of Rationality'. *The Philosophical Quarterly* 61, pp. 392–9.
—— 2014. 'The Real Symmetry Problem(s) for Wide-Scope Accounts of Rationality'. *Philosophical Studies* 170, pp. 443–64.
MacAskill, William ms. 'Meta Decision Theory and Newcomb's Problem'.
MacFarlane, John 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.
Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.
Parfit, Derek 2011. *On What Matters*. Vol. 1. Oxford: Oxford University Press.
Regan, Donald 1980. *Utilitarianism and Co-Operation*. Oxford: Oxford University Press.
Sartre, Jean-Paul 1946/2007. *Existentialism is a Humanism* (trans. Macomber). New Haven: Yale University Press.
Schroeder, Mark 2004. 'The Scope of Instrumental Reason'. *Philosophical Perspectives* 18, pp. 337–64.
—— 2009. 'Means-End Coherence, Stringency, and Subjective Reasons', *Philosophical Studies* 143, pp. 223–48.
—— 2010. 'Ought, Agents, and Actions'. *Philosophical Review* 120, pp. 1–41.
Sepielli, Andrew 2009. 'What to Do When You Don't Know What to Do'. *Oxford Studies in Metaethics* 4, pp. 5–28.

 © Pittard and Worsnip 2016

——2014. 'What to Do When You Don't Know What to Do When You Don't Know What to Do...'. *Noûs* 48, pp. 521–44.
Setiya, Kieran 2004. 'Against Internalism'. *Noûs* 38, pp. 266–98.
Silk, Alex 2014. 'Why "Ought" Detaches: Or, Why You Ought to Get with My Friends (If You Want to Be My Lover)'. *Philosophers' Imprint* 14.
Snedegar, Justin 2013. 'Contrastive Semantics for Deontic Modals'. In *Contrastivism in Philosophy*, edited by Martijn Blaauw, pp. 116–33. New York: Routledge.
Titelbaum, Michael G. 2015. 'Rationality's Fixed Point (Or: In Defense of Right Reason)'. *Oxford Studies in Epistemology* 5, pp. 253–94.
Way, Jonathan 2010. 'Defending the Wide-Scope Approach to Instrumental Reason'. *Philosophical Studies* 147, pp. 213–33.
——2011. 'The Symmetry of Rational Requirements'. *Philosophical Studies* 155, pp. 227–39.
Weatherson, Brian 2014. 'Running Risks Morally'. *Philosophical Studies* 167, pp. 141–63.
Wedgwood, Ralph 2016. 'Objective and Subjective "Ought"'. In Nate Charlow and Matthew Chrisman, eds., *Deontic Modality*, pp. 143–68. Oxford: Oxford University Press.
——2006. 'The Meaning of "Ought"'. *Oxford Studies in Metaethics* 1, pp. 127–60.
Wells, Ian ms. 'Known Divergence'.
Worsnip, Alex 2015a. 'Narrow-Scoping for Wide-Scopers'. *Synthese* 192, pp. 2617–46.
——2015b. 'Rationality's Demands on Belief'. Doctoral dissertation, Yale University.

 © Pittard and Worsnip 2016

Truth is Simple

Leon Horsten
University of Bristol
leon.horsten@bristol.ac.uk

Graham E. Leigh
University of Gothenburg
graham.leigh@gu.se

Even though disquotationalism is not correct as it is usually formulated, a deep insight lies behind it. Specifically, it can be argued that, modulo implicit commitment to reflection principles, all there is to the notion of truth is given by a simple, natural collection of truth-biconditionals.

> In contrast to an arbitrary procedure for moving from A_k to A_{k+1}, a reflection principle provides that the axioms of A_{k+1} shall express a certain trust in the system of axioms A_k.
>
> Solomon Feferman (1962, p. 261)

1. Introduction

John Burgess published a paper with the title 'The Truth is Never Simple' (Burgess 1986). What he meant was that the *extension* of the truth predicate in a typed, and even more so in a type-free approach, is complicated. This cannot be disputed. But we argue that the *intension* of the truth predicate is simple, in the sense that the content of the concept of truth is given by a simple and natural collection of truth-biconditionals. In other words, we claim that some form of *disquotationalism* must be in some sense correct. From a logical point of view, this takes us to the area of proof-theoretic approaches to truth, and away from the area of model-theoretic approaches to truth, which was the focus of Burgess (1986).

Arguments by Shapiro (1998) and Ketland (1999), based on observations by Tarski, have shown that certain standard formulations of disquotationalism are untenable. The fact that truth is compositional cannot be fully accounted for by disquotational axioms alone. Moreover, disquotational principles alone do not seem to do *justice* to the role that truth plays in metamathematical reasoning. In particular, compositional truth principles can be used to show that

Mind, Vol. 126 . 501 . January 2017 **© Horsten and Leigh 2016**
doi:10.1093/mind/fzv184 Advance Access publication 8 November 2016

reflection principles hold, and thus to justify reflection principles, whereas disquotational principles are too weak to do this.

Our position in this article is that disquotational principles none the less capture the core content of the concept of truth. When reflection principles are applied to (proof-theoretically weak) disquotational principles against the background of a weak syntax theory, strong compositional theories result. And when we are committed to a weak disquotational theory of truth, then we are *implicitly committed* to reflection principles for it. Therefore the compositionality of truth is implicitly contained in disquotational principles.

Shapiro and Ketland argue that reflection principles are justified by appeal to compositional truth principles. This road is patently not open to us. So the onus is on us to deliver an alternative account of our epistemic warrant for reflection principles. We shall provide this by appealing to Tyler Burge's distinction between *justification* for and *entitlement* to beliefs. We argue that we are commonly in a situation where we are entitled to rely on and even believe in reflection principles, even though we do not have a justification for them.

The structure of this article is straightforward. We first discuss traditional forms of disquotationalism. Then we turn to the critique of disquotationalism by Shapiro and Ketland. Subsequently we outline how compositionality follows from disquotational theories, modulo reflection principles. (Sketches of proofs are relegated to a technical appendix.) We then compare our view with proposals by Field and Halbach that seek somehow to derive the compositional nature of truth from disquotational principles. Finally, we give an epistemological account of the notion of implicit commitment to reflection principles.

In the literature, the discussion of the relation between disquotationalism and reflection principles is mostly restricted to a typed setting, where Tarski's distinction between metalanguage (the language of truth) and object language (the language of the background syntax theory, which we identify with a weak arithmetical theory) is maintained. In this article, we shall not only be occupied with typed theories, but will give an account of the relation between disquotationalism and compositionality in a type-free environment as well.

We assume that the reader is familiar with the basic theories of arithmetic *EA* (Elementary Arithmetic) and *PA* (Peano Arithmetic).[1]

[1] For the purposes of this paper we assume *PA* is formulated with an additional function symbol 2^x for binary exponentiation and the axioms $2^0 = 1$ and $\forall x(2^{x+1} = 2^x + 2^x)$; *EA* is then the sub-theory of *PA* in which induction is restricted to bounded (Δ_0) formulae.

 © Horsten and Leigh 2016

These will be used as background theories of syntax, modulo coding. In the interest of readability, we shall be somewhat sloppy with the details of coding (except in the technical appendix), and generally identify a formula ϕ with its code $\ulcorner\phi\urcorner$. We also assume familiarity with the most important typed and type-free disquotational and compositional theories of truth. Specifically, we assume that the reader is acquainted with *TB* (Tarski-biconditionals), *UTB* (*uniform* Tarski-biconditionals), *CT* (Compositional Truth), and *KF* (Kripke-Feferman). Precise definitions and extended discussions of these theories can be found in the appendix at the end of this paper and in Halbach (2011).

Reflection principles play a central role in this article. In particular, we shall be concerned with:

(1) *Local reflection principles* (Rfn_S), which are schematic principles of the form $Prov_S(\phi) \rightarrow \phi$, where ϕ ranges over a collection of sentences of some language $\mathcal{L}$, S is a theory in $\mathcal{L}$, and $Prov_S$ is a canonical provability predicate for S.

(2) *Uniform reflection principles* (RFN_S), which are schematic principles of the form $\forall x : Prov_S(\phi(x)) \rightarrow \phi(x)$, where $\phi(x)$ now ranges over a collection of formulae with one free variable of some language.

(3) *Global reflection principles* (GRP_S), which are axioms of the form

$$\forall \phi \in \mathcal{L} : Prov_S(\phi) \rightarrow T\phi$$

For a brief discussion of these types of reflection principles, see Halbach (2011, ch. 22).

2. From Tarski to disquotationalism

Tarski (1935, pp. 187ff.) famously proposed *Convention T* as a (material) adequacy condition on a definition of truth for a language $\mathcal{L}$:

> A truth definition for $\mathcal{L}$ must entail, for every sentence $\phi \in \mathcal{L}$, the *Tarski biconditional*
>
> 'ϕ' is true $\leftrightarrow \phi$

Tarski's theorem of the undefinability of truth entails that a sufficiently strong theory cannot consistently contain a truth definition for its own language that satisfies Convention T. But natural language

 © Horsten and Leigh 2016

seems maximally expressive, in the sense that whatever can be expressed at all can be expressed in natural language. So no matter how strong a consistent theory we formulate in natural language, it cannot contain a truth predicate that satisfies Convention T. Yet natural language does contain a concept of truth.

Tarski concluded from this that the concept of truth in natural language must therefore be incoherent (Tarski, 1935, p. 264). But at the same time it is clear that there are extensive fragments of natural language, and scientific theories that do not contain the concept of truth, that are coherent. For such languages *cum* theories, a truth definition can be constructed in a more expansive metalanguage: Tarski showed us how.

In the 1950s and 1960s, under the influence of Wittgenstein and his *ordinary language* philosophy, natural language was to some extent rehabilitated in analytical philosophy (Soames 2003). This rehabilitation extended to the concept of truth. Moreover, ordinary language philosophers drew attention away from precise extensions of concepts and instead emphasized the *use* of concepts in ordinary language. Applied to the concept of truth, this means that, *pace* Tarski, giving a definition of the extension of the concept of truth in natural language is not called for. We should try to capture the use of the concept of truth without trying to define truth.

Tarski's Convention T provides key information about the use of the concept of truth. It is a device for *quotation* (right-to-left) and for *disquotation* (left-to-right). If this is indeed the central (and perhaps the only) function of the concept of truth, then its use can be captured in an infinite collection of *axioms*. They are called *truth-biconditionals* (or *Tarski-biconditionals*).

Ultimately, we are interested in providing a satisfactory theory of truth for natural language. But, as is customary in truth theory today, in order to test disquotationalism (and ideas like it), we will work with a toy language, a miniature version of natural language. The miniature language that we opt for is $\mathcal{L}_T$, which consists of the language of arithmetic $\mathcal{L}_{PA}$ plus a primitive truth predicate (T).

Suppose now that we want to describe the use of the concept of truth for $\mathcal{L}_T$. Then, inspired by Convention T, and determined to keep metalanguage and object language separate, we can suggest the restricted truth-biconditionals for this purpose:

 © Horsten and Leigh 2016

Axiom 1 (TB) For all sentences $\phi \in \mathcal{L}_{PA}$: $T(\phi) \leftrightarrow \phi$

This axiom scheme to a significant extent captures the disquotational role of the concept of truth.

As was mentioned in the introduction, TB presupposes a background theory of syntax. Ultimately, we will opt for a very weak background theory, namely, Elementary Arithmetic (*EA*).[2] But more often than not, in the literature, a stronger background is presupposed, namely, Peano Arithmetic. For the time being, *TB* will denote the restricted Tarski-biconditionals with an unspecified background arithmetical theory, which can be *EA*, *PA* without the truth predicate appearing in the induction scheme (henceforth simply *PA*), or *PA* with induction extended to the language including the truth predicate (PA_T). When choice of background theory matters, we will say so and be more definite. In particular, later in this article we will officially take *EA* as our background syntax theory. In fact, one of the aims of the present article is to sidestep discussions about the background syntax theory.[3]

Some authors believe that free paramaters should be allowed in the disquotational axiom. This yields the stronger axiom scheme:

Axiom 2 (UTB). For all formulae $\phi(\vec{x}) \in \mathcal{L}_{PA}$: $\forall\vec{x}[T(\phi(\vec{x})) \leftrightarrow \phi(\vec{x})]$

Again, without risk of confusion we can denote by *UTB* the theory resulting from adding axiom 2 to a background syntax theory.

It is not hard to see that axiom 2 entails the distribution of truth over the universal quantifier: $\forall x T(\phi(x)) \leftrightarrow T(\forall x \phi(x))$. The truth of this statement is directly supported by an intuition that says that truth is compositional. But axiom 2 somewhat exceeds the immediate content of disquotationalism. (Indeed, it turns out that the theory *TB* does not entail axiom 2.) So we will take the content of the disquotational intuition to be captured by a scheme that quantifies over *sentences* rather than over formulae with free variables.

At first sight, it appears that the theory *TB* is *incomplete*: the restricted truth-biconditionals say nothing about the disquotational role of sentences that themselves contain the concept of truth.[4] So, *a fortiori*, the restricted Tarski-biconditionals cannot be seen as a *definition* of truth for $\mathcal{L}_T$. But Tarskian considerations suggest that this may be a

[2] *EA* is defined in the appendix, definition 1.

[3] More about this below: see §5.

[4] As mentioned earlier, later in this article we will also consider truth theories that speak to the disquotational role of sentences that themselves contain the concept of truth.

 © Horsten and Leigh 2016

virtue rather than a vice. Tarski's proof of the undefinability of truth is based on instantiating the liar sentence into the Tarski-biconditional scheme. But the liar sentence itself contains the concept of truth. In sum, according to Tarski's diagnosis of the liar paradox, we are well advised not to allow the concept of truth inside the truth-biconditionals. This is known in the literature as *typing* the concept of truth. And this is of course precisely what happens in TB.

We thus arrive at *TB* as a natural disquotational theory of truth. Disquotationalists hold that quotation and disquotation are the only functions of the concept of truth: the (restricted) Tarski-biconditionals are all there is to truth (Williams 1988, p. 424). Quine put it as follows:

> The truth predicate is a device of disquotation. We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences, then the truth predicate has its use. (Quine 1970, p. 307)

A version of disquotationalism in terms of propositions instead of sentences is defended in Horwich (1998). He emphasizes that truth is a *simple concept*:

> [The Tarski-biconditionals] could be explained only by principles that are simpler and more unified than they are—principles concerning propositional elements and the conditions in which truth emerges from combining them. But the single respect in which the body of minimal axioms is not already perfectly simple is that there are so many of them—infinitely many; and no alleged explication could improve on this feature. For there are infinitely many constituents to take into account: so any characterization of them will also need infinitely many axioms. (Horwich 1998, p. 51)

We shall argue that in this Horwich is basically correct: the core content of the concept of truth is captured by a natural and simple collection of Tarski-biconditionals.

3. Compositionality and the limitations of disquotationalism

There is a fly in the ointment—as Tarski already knew. It seems fundamental to our concept of truth that it is a *compositional* notion: the concept of truth distributes over the logical symbols. For instance, it seems that the following statement is acceptable:

$$\forall \phi, \psi \in \mathcal{L}_{PA} : T(\phi \wedge \psi) \leftrightarrow [T(\phi) \wedge T(\psi)]$$

 © Horsten and Leigh 2016

Tarski rejected *TB* as a theory of truth because of its deductive weakness:

> The value of [the theorem that *TB* is consistent] is considerably diminished by the fact that [the Tarski-biconditionals] have a very restricted deductive power. A theory of truth founded on them would be a highly incomplete system, which would lack the most important and most fruitful general theorems. (Tarski 1935, p. 257)

Indeed, it is not hard to see that *TB* does not entail compositional truth principles such as principle (1). A proof from *TB* (or *UTB*) can only use finitely many Tarski-biconditionals, whilst the compositional principles make a claim about infinitely many sentences. Every instance of the aforementioned compositional principles can be proved from *TB*; the universal closure cannot. This is why *TB* is incomplete as a theory of truth.

Moreover, there are statements from the background theory *S* that we expect a theory of truth for *S* to prove. For a sufficiently strong background theory *S*, we expect a theory of truth to prove the Gödel sentence G_S of *S* (Shapiro 1998; Ketland 1999, §6) and the local reflection principle Rfn_S for *S*. When we set set $S \equiv PA$, and take *PA* as our background arithmetical theory, we should expect *TB* to prove G_{PA} and Rfn_{PA}, but it does not, because *TB* is proof-theoretically conservative over *PA* (Halbach 2011, ch. 7).

We expect a truth theory for *PA* to prove G_{PA} and Rfn_{PA} because their truth can be established by *truth-theoretic reasoning*. Let us consider Rfn_{PA} first. The axioms of *PA* and of logic are all true. The rules of inference preserve truth. Therefore every theorem of *PA* is true. Now consider G_{PA}. By the diagonal lemma, this sentence is true if and only if it is unprovable in *PA*. Suppose G_{PA} were provable in *PA*. Then, since Rfn_{PA} holds, G_{PA} would be true. But by the diagonal lemma that means that it would be unprovable in *PA*. Contradiction. So G_{PA} is unprovable in *PA*. So by the diagonal lemma again, G_{PA} is true.

From this, truth theorists tend to draw the conclusion that the compositionality intuition about truth is not reducible to the disquotational intuition (Horsten 2011, p. 70). An option that is taken by many at this point is to take the compositional truth principles as *basic* axioms of truth. This option is taken in, for instance, Shapiro (1998) and Ketland (1999). This leads to the proposal of accepting, instead of *TB*, the compositional theory *CT* (which includes principles such as (1) as *axioms*) as a basic theory of truth.

 © Horsten and Leigh 2016

The advantage of this proposal is threefold. First, *CT* is a simple and natural set of axioms for truth. As a truth theory, it is not as simple as *TB*, but it is still a simple theory of truth. That it is natural is shown by the fact that it has been embraced by leading philosophers such as Davidson (1967). Second, *TB* is a sub-theory of *CT* (Horsten 2011, ch. 6). So the disquotational intuition is contained in the compositional intuition as expressed by *CT*. Third, *CT* does entail meta-mathematical statements such as G_{PA} and Rfn_{PA} that can be recognized to be true on the basis of truth-theoretic reasoning (Ketland 1999).

The structure of the standard proof of the non-conservativeness of *CT* goes as follows. First, by an induction, in *CT*, on the length of proofs in *PA*, the *global reflection principle* GRP_{PA} for *PA* is proved. Then, in a second step, the truth-biconditionals for *PA* (which are provable in *CT*) are used to derive the local reflection principle Rfn_{PA}, which is of course unprovable in *PA*.

It has recently become clear that the moral of our discussion also extends to the type-free setting. In particular, it applies to Feferman's system *KF*, which is the most popular type-free compositional theory of truth.[5] In particular, is commonly recognized to be a very natural way of extending *CT* to a consistent type-free system.

The unrestricted truth-biconditionals are of course inconsistent. Until recently, it was not clear how to construct a natural consistent type-free disquotational theory of truth. However, the situation changed with the publication of Halbach (2009). In this article, Halbach proposes the *Positive Uniform Tarski-biconditionals* (PUTB), containing the universal closure of every truth-biconditional $T(\phi(\vec{x})) \leftrightarrow \phi(\vec{x})$, where ϕ is a formula of $\mathcal{L}_T$ in which the truth predicate only occurs in the scope of an even number of negation signs: call such formulae *truth-positive formulae*. Halbach proves that PA_T + PUTB is a sub-theory of *KF* and can define a truth predicate that satisfies the *KF* axioms.

As before in the typed setting,[6] the scheme PUTB exceeds the content of disquotationalism somewhat. So, for now, we will take as a natural candidate for a type-free disquotational scheme the following:

[5] Another popular, but decidedly less popular, type-free compositional theory of truth is the Friedman-Sheard system *FS* (Friedman and Sheard 1987). The main drawback of *FS* is that it is ω-inconsistent. For that reason, we leave it aside in the present article.

[6] See our comparison of *TB* with *UTB* in the previous section.

 © Horsten and Leigh 2016

Axiom 3 (PTB). For all truth-positive sentences $\phi \in \mathcal{L}_T$: $T(\phi) \leftrightarrow \phi$

The result of adding axiom 3 to the background theory is called *PTB*, and the extension by the uniform version of axiom 3 is called *PUTB*. *PTB* is a weak truth theory: it is conservative over its background theory (Cieśliński 2011).[7] For the proof-theoretic strength of *PUTB* it matters whether induction is permitted for the truth predicate. If induction is not extended, then it is proof-theoretically conservative over the background theory; if induction on all formulae of $\mathcal{L}_T$ is chosen, then it is highly non-conservative over the background theory. Again, we need not be very concerned with this phenomenon in this paper for reasons that will be explained later. At any rate, neither *PTB* nor *PUTB* proves the compositional principles of truth, not even the typed ones (see, for example, Halbach 2011, lemma 19.20). The argument is structurally identical to the proof that *TB* and *UTB* cannot prove the compositional axioms of *CT*. So in the type-free setting too, the compositionality of truth seems to have more content than the disquotationality of truth.

4. Reflection and justification

Let us return to the typed setting. In §3 the conclusion was drawn that *TB* is too weak and should instead be replaced by *CT* (Shapiro 1998; Ketland 1999). Tennant (2002) takes it upon himself to defend the disquotationalist against this line of reasoning.

Tennant first points out that it is not the case that metamathematical justification crumbles if one is only allowed to make use of a conservative, disquotational concept of truth. For instance, it is a commonplace that one can prove G_{PA} from weak reflection principles such as Rfn_{PA}.[8] Secondly, weak reflection principles express '[one's] willingness …, via [Rfn_{PA}], to assert any theorem of [*PA*]' (Tennant 2002, p. 574).

Ketland, in his reply to Tennant, expresses dissatisfaction with Tennant's strategy for simply *postulating* reflection principles. He replies that reflection principles such as Rfn_{PA} ought instead to be *proved* (Ketland 2005, p. 82), and stresses that the truth theory *CT* does indeed prove such reflection principles. In other words, Ketland objects that Tennant's justification of G_{PA} is shallower than his own

[7] See also theorem 12 in the appendix

[8] In fact, Tennant stresses that in order to prove G_{PA} the full force of Rfn_{PA} is not needed. All that is needed is the schematic principle that asserts $Prov_{PA}(\phi) \rightarrow \phi$ for every primitive recursive formula ϕ of the language of *PA*.

 © Horsten and Leigh 2016

justification, for some of the justifying principles (viz., the reflection axioms) that are postulated in Tennant's justification are themselves proved in Ketland's justification of G_{PA}.

Tennant rejects Ketland's request for a justification of reflection principles such as Rfn_{PA}. He writes:

> No further justification is needed for the new commitment made by expressing one's earlier commitments. As soon as one appreciates the process of reflection, and how its outcome is expressed by the reflection principle, one already has an explanation of why someone who accepts S should also accept all instances of the reflection principle. (Tennant 2005, p. 92)

Ketland finds this response unsatisfactory. He asks, first, what the reflective process of which the passage speaks is supposed to be. Tennant is not explicit about this, so Ketland offers an answer to this question. The reflective process consists in stepping back from one's practice and realizing that one is ready to accept every theorem of *PA* (Ketland 2010, p. 428). So the outcome of the reflection process, in Ketland's view, is the conclusion 'I am ready to accept every theorem of *PA*' (Ketland 2010, p. 433). Second, Ketland disputes that the realization of this disposition is adequately expressed by Rfn_{PA}:

> It should be noted that this is a non-standard claim. Usually, a reflection scheme like $[\mathrm{Rfn}_{PA}]$ is said to express the *soundness* of $[PA]$: that whatever $[PA]$ proves is *true*. And being *true* is not the same as being accepted. (Ketland 2010, p. 430)

All this underscores that the process of reflection is not well understood: it is not clear what it consists in, what its outcome is, and how that can be formally expressed (if it can be formally expressed at all). Ketland holds that reflection principles should be and can be justified by means of principles of truth, whereas Tennant regards reflection principles as the outcomes of reflective processes that need no further justification.

Cieśliński then enters the debate between Ketland and Tennant (Cieśliński 2010). First of all, he takes CT^-, the version of *CT* which has *PA* (induction only for $\mathcal{L}_{PA}$ formulae) as a background theory but where the truth predicate is not allowed in the induction scheme, as an unobjectionable base theory. The theory CT^- goes beyond *TB* in one sense, because *TB* cannot prove the compositional truth axioms. Thus CT^- cannot be seen as an expression of disquotationalism. But CT^- is weaker than *TB* in another sense: the truth predicate is not allowed in the induction scheme of CT^-, and hence the theory is proof-theoretically conservative over its background

 © Horsten and Leigh 2016

theory (Kotlarski et al. 1981).[9] The background theory of *TB* may, however, feature the full induction scheme missing from CT^-. So in that sense, many deflationists who are not strict disquotationalists find CT^- unproblematic.

Cieśliński agrees with Tennant (contra Ketland) that reflection principles are not themselves in need of justification. He furthermore holds that *uniform* reflection principles express the outcome of reflective processes. But he agrees with Ketland that even *global* reflection principles, which allow us to express the content of an infinite sequence of uniform reflection principles, ought to be provable.

Cieśliński shows that $CT^- + \mathrm{GRP}_{PA_0}$ does not prove RFN_{PA}, where we take PA_0 to stand for Peano Arithmetic formulated in the language *without* the truth predicate. In other words, adding a strong reflection principle for the *restricted* language $\mathcal{L}_{PA}$ to CT^- does not give us even uniform reflection for the *extended* language $\mathcal{L}_T$. This means that adding reflection principles for the restricted language cannot give us what we want.

Cieśliński considers the following uniform reflection principle:

Axiom 4 ($\mathrm{RFN}_{\varnothing}$) $\forall x : Prov_{\varnothing}(\phi(x)) \rightarrow \phi(x)$ for all formulae $\phi \in \mathcal{L}_T$

where $Prov_{\varnothing}$ expresses provability from the empty theory, that is, first-order logical theoremhood. So this principle expresses the truth of theorems of (first-order) logic in the extended language.

Then Cieśliński shows that in CT^-, the full induction axiom for the extended language $\mathcal{L}_T$ follows from axiom 4 (Cieśliński, 2010, p. 419):

Theorem 1. $CT^- + \mathrm{RFN}_{\varnothing} \vdash CT$

But we know that CT entails the global reflection principle for the extended language $\mathcal{L}_T$, which is all the reflection that one might want in a typed setting.

In the type-free setting, something similar can be said. The Kripke-Feferman theory without truth allowed in the induction scheme, often called KF^-, is conservative over its background theory *PA*. But if we add reflection over logic to KF^- in the extended language, then we obtain the full (and highly non-conservative) theory *KF* (Cieśliński, 2010, p. 421).

Now Cieśliński argues that axiom 4 is a reflection principle that is the outcome of a reflective process of the sort that Tennant describes:

[9] CT^- is not *semantically* conservative over its background theory (Lachlan 1981): not every model of arithmetic can be expanded to a model of CT^-. Neither is *TB* in the case that its background theory contains induction for the truth predicate (Strollo 2013).

 © Horsten and Leigh 2016

if we reflect upon on our inferential practices, we see that we accept the laws of logic even in the extended language. Note that axiom 4, *unlike* Rfn_{PA}, is a uniform reflection principle. If one agrees with Ketland that the outcome of reflection is not captured by a local reflection principle, then one will *a fortiori* deny that it is captured by a uniform reflection principle. On the other hand, if one believes that the outcome of a reflection process is captured by a reflection principle, then it seems that a uniform reflection principle is as good a candidate as a local reflection principle. Also, even though it is essential that axiom 4 quantifies over all sentences in the extended language $\mathcal{L}$, the truth predicate does not play an 'active' role in this principle. In other words, truth does not play a *substantial* role in axiom 4.

The key element in Cieśliński's account is the suggestion that we can not only reflect on a theory in our background language, but that we can also reflect on our background theory in an *extended language*. Indeed, when we commit ourselves to logic, or to the principle of induction, we commit ourselves to open-ended schemes. In Feferman's terms, logic and Peano arithmetic are *schematic* theories (Feferman 1991, p. 2). Whenever a new bona fide predicate enters our language, we automatically extend these schemes to the new language.

In Cieśliński's process of passing from CT^- to CT via a reflection on logic, conservativeness over the background theory is lost. So from the point of view of a deflationist who holds that a theory of truth ought to be proof-theoretically conservative, the theory $CT^- + \mathrm{RFN}_{\varnothing}$ does not seem attractive as a basic truth theory. But it ought to be remembered that it is not proposed as a *basic* truth theory. The basic truth theory is the conservative theory CT^-. But, especially if, as Field once suggested, truth is a *logical* notion (Field 1999, p. 534), we seem to be implicitly committed to axiom 4.

So that means that, given the basic truth theory CT^-, we are implicitly committed to the full theory CT, and thus to the global reflection principle for arithmetic. In other words, modulo reflection on logic, truth can perform the metamathematical functions (for instance, justifying G_{PA}) that it is meant to.

5. Reflection and disquotation

A key aspect of Cieśliński's proposal is the following. Whereas Ketland proves reflection principles from truth principles, and Tennant argues

 © Horsten and Leigh 2016

that reflection principles need no 'proof', Cieśliński reverses the explanatory direction, at least to some extent. Reflection principles allow us to derive a stronger truth theory (CT) from a weaker one (CT^-). In what follows, we will take Cieśliński's strategy much further. We will show, in fact, that strong compositional truth theories follow from local truth-biconditionals, via reflection principles.

According to the strict disquotationalist, our basic truth theory is *TB*. As far as the truth axioms go, *TB* is weaker than CT^-, which is taken by Cieśliński as the basic truth theory. But as far as the arithmetical principles go, *TB* may be stronger, that is, if it allows the truth predicate in the induction scheme. But, as we have argued earlier, it is reasonable to allow this, for Peano arithmetic is a *schematic* theory. Indeed, CT^- does not seem to represent a stable position. It extends the logical schemes so as to allow occurrences of the truth predicate, but does not extend the same privilege to the induction scheme.

Nevertheless, as announced earlier, we will sidestep the discussion about the background syntax theory and start with an arithmetical theory that does not contain strong schematic principles: Elementary Arithmetic (*EA*).[10] So from now on, we let TB denote the Tarski-biconditionals for the language of arithmetic, and fix *EA* as the background theory. Then our official disquotational starting point is $TB_0 = EA + \mathrm{TB}$.[11]

Cieśliński's use of implicit commitment is very modest: it is restricted to the reliability of *logic*. But implicit commitment is a general notion. Whenever we have fully committed ourselves to a theory *S*, we are implicitly committed to accepting a reflection prinicple for *S* (Feferman 1991, p. 1). Since the disquotationalist's starting point is TB_0, she is implicitly committed to a reflection principle for TB_0. We have argued earlier that the reflective commitment extends at least to *uniform* reflection over the starting point. That is, the disquotationalist is implicitly committed to RFN_{TB_0}.

[10] It is entirely possible to consider a base theory that is completely free of schematic principles. One option is simply to take as a base theory $I\Sigma_1$, the extension of *EA* by induction for Σ^0_1 formulae, which is known to permit a finite axiomatization. If $I\Sigma_1$ is considered too strong as a background theory, it is possible to proceed using a finitely axiomatized sub-theory of *EA*: cf. footnote 9.2 below.

[11] We adopt the convention in this paper that theories notated using subscripts are formulated without induction for the extended language. Theories notated without subscripts, such as *TB*, *CT*, *KF*, etc., always contain induction for the whole language.

 © Horsten and Leigh 2016

The disquotationalist can reflectively come to realize that she is implicitly committed to RFN_{TB_0} and explicitly accept her commitment to RFN_{TB_0}. Then she has accepted the stronger theory

$$EA + \mathrm{RFN}_{TB_0}$$

Let us call this theory TB_1. In the process, she has also extended her implicit commitment to $\mathrm{RFN}_{EA+\mathrm{RFN}_{TB_0}}$. She can reflectively come to realize this, and explicitly accept

$$EA + \mathrm{RFN}_{EA+\mathrm{RFN}_{TB_0}}$$

We shall call this theory TB_2. It can then be said that a commitment to TB_2 is implicit ('implicitly implicit') when one explicitly signs up to TB. This process TB_0, TB_1, TB_2, … can justifiably be iterated along a transfinite ordinal hierarchy that is as long as the length of verifiable well-orderings that can be generated in the process, in the spirit of Feferman (1991). This transfinite process results in a full description of the implicit commitment of *TB*, which is called the *reflexive closure* of TB. Indeed, in a similar fashion, reflection hierarchies of other truth theories can be constructed, such as $EA + \mathrm{UTB} = UTB_0$, UTB_1, UTB_2, … But for the purposes of this article, we shall be interested only in the two first movements of extending the basic truth theory by reflection principles.

It can be shown that:[12]

Theorem 2. $TB_1 \vdash UTB + \mathrm{Ind}(\mathcal{L}_T)$

where $\mathrm{Ind}(\mathcal{L}_T)$ is the induction scheme for the language $\mathcal{L}_T$.

So in the first reflective moment, the truth theory is strengthened from the local Tarski-biconditionals to the uniform version. And someone who accepts the disquotational axioms is at least implicitly committed to induction in the extended language. When a person *reflects* on this commitment implicit in *TB*, she explicitly accepts Peano Arithmetic in the extended language. In other words, for a reflective person, the discussion whether truth should be allowed in the induction scheme is immaterial. This is how we sidestep the discussion in the literature on the appropriate background syntax theory for an axiomatic truth theory.

[12] See appendix.

 © Horsten and Leigh 2016

Moreover, it has also been shown (Halbach 2001, §4) that:

Theorem 3. *CT is identical to* UTB_1 *and a sub-theory of* TB_2.

This means that in the second reflective moment, the full compositionality of truth is obtained. In other words, the compositionality of truth is implicitly contained in the disquotational axioms. The compositional axioms for conjunction, disjunction and negation are obtained after just one act of reflection on the local biconditionals (that is, in TB_1); a second instance of reflection is necessary in order to derive the compositional quantifier axioms.

Let us now turn to the type-free environment. Here the positive Tarski-biconditionals (PTB) seem a possible starting point for the disquotationalist. From the compositional point of view, *KF* seems a natural theory. Again, we can consider the hierarchy $PTB_0 = EA + \text{PTB}$, PTB_1, $PTB_2, \ldots$ of truth theories generated from the positive truth-biconditionals by repeated application of uniform reflection. Then we have

Theorem 4. (Halbach 2009). *KF is interpretable in* PTB_1.

Theorem 5. (Halbach 2001). *The 'strictly positive' compositional axioms of KF are derivable in* PTB_2.

Theorem 4 has to be interpreted with care. Halbach's proof (see the appendix for an outline) shows that *KF* (in any reasonable formulation) can be interpreted in a very *simple* manner into *PUTB*, a subtheory of PTB_1. As in the type-free scenario, it is only after two instances of reflection that the compositional axioms for quantifiers become provable from local biconditionals. Nonetheless, *KF* (in any formulation) is not a sub-theory of PTB_2. For example, in the formulation of *KF* with a single truth predicate, the compositional axioms involving negation (e.g. $T(\neg T(\psi)) \leftrightarrow T(\neg \psi)$) will not derivable. So we refrain from claiming that someone who accepts PTB_0 is implicitly committed to the compositional theory *KF*.

McGee (1992) has shown, using a diagonal argument, that *every* theory in $\mathcal{L}_T$ can be written as a collection of Tarski-biconditionals in $\mathcal{L}_T$. But the collection consisting of the positive Tarski-biconditionals forms a *natural* theory. Moreover, we know that it is consistent, because it forms a sub-theory of *KF* (Halbach 2009). Nonetheless, it would be desirable to have an argument that *motivates* taking the positive Tarski-biconditionals as one's basic disquotational truth axioms in a type-free setting. In keeping with the main tenets of

 © Horsten and Leigh 2016

disquotationalism, we want such a motivation to start from a basic collection of Tarski-biconditional sentences.

Consider the language $\mathcal{L}_P$ (where '*P*' is short for 'partial'), which is just like the language $\mathcal{L}_T$ except that:

(1) for every atomic predicate A, there is an atomic dual $\bar{A}$ of A;

(2) the negation symbol is removed from the language.

Let the dual of T be denoted as F. Aside from the unrestricted truth-biconditionals for $\mathcal{L}_P$, we also consider the *falsity-biconditionals* for this language, which are the sentences of the form $F(\bar{\phi}) \leftrightarrow \phi$, where $\bar{\phi}$ is the result of replacing in ϕ each connective, quantifier and atomic predicate by its dual.[13] Consider the theory TFB_0 that consists of EA plus the unrestricted truth- and falsity-biconditionals in $\mathcal{L}_P$. The theory TFB_0 seems a good basic starting point. It is no more than a totally unrestricted collection of truth-biconditionals and falsity-biconditionals in a language without negation. In particular, PTB_0 is a sub-theory of TFB_0. Here the concept of negation is seen as the source, when combined with the truth predicate, of the semantic paradoxes.

Then it can be shown that:[14]

Theorem 6. $TFB_1 = EA + \mathrm{RFN}_{TFB_0} \vdash \mathrm{PUTB}$

Theorem 7. *KF is a sub-theory of* $UTFB_1$ *and is strictly contained in* TFB_2.

Here the provability relation $\vdash$ denotes (as always in this article) provability in *classical* predicate logic.

If we string these results together, then we see that in the type-free setting, the full compositional axioms of *KF* can be seen to be implicit (via two reflective moments) in the unrestricted truth- and falsity-biconditionals in a 'liar-proof', negation-less language.

In sum, both in the typed and in the untyped setting, the compositional content of the concept of truth can be taken to derive, through reflection principles, from a purely disquotational conception of truth. So perhaps truth is a very simple concept after all.

Of course, all of this leaves one fundamental question unanswered: how are reflection principles justified? This question was forcefully put on the agenda by Ketland (2005, p. 85). Indeed, it is surprising that

[13] For the precise definition of $\bar{\phi}$, see the appendix.

[14] See appendix.

 © Horsten and Leigh 2016

even though hierarchies of reflection principles had been studied by Kreisel and Feferman since the 1950s, and that they suggested early on that we are implicitly committed to reflection principles for the theories that we accept,[15] philosophers of mathematics have hitherto largely failed to investigate the notion of *implicit commitment*, and have not spent much philosophical energy on analysing our warrant for reflection principles.

We have seen that in Ketland's view, reflection principles are justified by appeal to (compositional) truth principles. Since we instead want to motivate compositional truth from reflection principles (and natural disquotational principles), this road is not open to us. Instead, we argue that our epistemic warrant underwriting reflection principles is of a different kind.

6. Reflection and entitlement

We will now argue that when we are justified in believing a theory, we do not need extra justification for adopting a reflection principle for that theory. In such a situation, we are entitled to adopt a reflection principle without giving additional justification for accepting it. In arguing our point, we draw upon the distinction between *entitlement* and *justification* that was made in Burge (2003).

Our claim goes against the majority view. We have seen that Ketland and Shapiro request a justification for reflection principles. This is the common viewpoint. Also Volker Halbach, for instance, requires justification in this situation: 'the transition from a theory to a reflection principle for that theory requires an argument' (Halbach 2001, p. 1963).

Burge has invoked the distinction between between justification and entitlement in his writings on perception, memory, self-knowledge and logical reasoning (Burge 1993, 1998, 2003, 2007). We believe that this distinction can do similar work in the epistemology of proof-theoretic reflection.

To provide background for our discussion, let us start with perception. We trust our senses. That we do so is made manifest in our everyday beliefs and actions. Most of us *believe* that we trust our senses, but we do not *need* to believe that we trust our senses in order to trust our senses.

[15] See, for instance, Kreisel (1970).

 © Horsten and Leigh 2016

Someone who has the required conceptual machinery can articulate our trust in our senses. Then it becomes a *reflection principle*:

What our senses tell us is (or tends to be) true.

Let us call this the *sense reflection principle*. We are *entitled* to believe in the trustworthiness of our senses. This entitlement does not require a justification of our belief in the trustworthiness of our senses. Indeed, perhaps the sceptic cannot be answered: perhaps there is no non-circular justification of our belief in the trustworthiness of our senses.

We can come to believe in the trustworthiness of our senses through a *reflective act*. When we do so, we acquire a priori belief in, and even knowledge of, the trustworthiness of our senses. The details of the reflective process are complicated. But, for our purposes, the important point is that this reflective process does not typically involve a justification of the trustworthiness of our senses. In sum, through a process of reflection we come to *know* the sense reflection principle. The latter is therefore not an axiom or basic principle, and it is also not justified on the basis of more basic principles.

Note also that we need not come to believe in the trustworthiness of our senses. After all, we do not normally *use* even instances of the sense reflection principle when we rely on our sensory experiences to form our beliefs about the world around us: we do not *reason* from our experiences to our beliefs.

Something similar can be said about our everyday reliance on preservative memory and our reliance on logical inference. We rely on memory in daily life, and we are usually entitled to do so without justification. When we remember something (in normal circumstances), we are entitled without further justification to self-attribute having a memory, but we can rely on our memory without doing so. When we reason, we rely on logical inference rules. Our normal entitlement of doing so is grounded in our knowledge of the logical constants (Burge 2007, p. 197).

There is an important difference between our reliance on perception and our reliance on logical reasoning. In the case of perception, there is always the possibility of *brute error*.[16] We may be in a situation where there is no malfunction of our perceptual faculties, yet we find ourselves in an inclement environment which causes our perceptual beliefs to go astray. There is no possibility of *brute* error in logical

[16] The question of the possibility of brute error in perception, self-knowledge, memory and logical reasoning is discussed in Burge (2007).

 © Horsten and Leigh 2016

reasoning (Burge 2007, p. 196). If our reasoning faculties function as they should, and are appropriately rooted in our understanding of the logical concepts, then we cannot go astray in our reasoning. This is not to say, of course, that faulty reasoning is not possible; it is just that faulty reasoning cannot simply be due to the world not cooperating.

Mistakes in perception, memory and logical reasoning do occur, and we know that they do. But this does not mean that we are invariably entitled only to less than full acceptance of what perception, our memory, or our logical faculties deliver. When our faculties do function appropriately (and, where relevant, the world cooperates), we are entitled to unqualified acceptance of the beliefs based on the functioning of those faculties: there is then no epistemic obligation to hedge.

If Burge's distinction, within the category of epistemic warrant, between justification and entitlement, is applicable to any other domain, then it must be applicable to the domain of reflection principles for mathematical theories and truth theories. In particular, we seek to apply this distinction to the *uniform* reflection principle for *TB*. And we have seen that we seek to apply uniform reflection at least *twice*. How are we to think of this?

Consider a congregation of believers attending mass in church. They accept what is written in the Gospels: they see it as the word of God and they trust God. They have not reflected on their acceptance of what is written in the Gospels. The priest, however, in his sermon articulates the community's trust in God; he reflects on the doxastic mode of the congregation. The members of the congregation could perhaps not even do this by themselves: before the priest has articulated it for them, they may lack the necessary conceptual machinery. But the congregation simply extends its trust to the priest, and (if all is well) it is entitled to do so, without proof or argument. And the priest can even reflect on this extended trust in his sermon …

So it is with reflection principles for mathematical theories and for truth theories. We trust basic arithmetic (*EA*), classical logic, and a basic collection of truth-biconditionals, perhaps even including a set of falsity-biconditionals. This is evident from the way in which we use what we establish in basic arithmetic and in basic disquotational theories. Just as our perceptual states (as representational states) are integrated into our belief system (Burge 2003, p. 521), so are our *arithmetical proof states* integrated into our belief system. Indeed, we indispensably use arithmetical theorems in our best explanations of physical phenomena. For instance, the fact that cicadas have

 © Horsten and Leigh 2016

prime-numbered periodical life-cycles (of thirteen or seventeen years) has been given an evolutionary explanation involving general facts about prime numbers (Baker 2005). Similarly, disquotational reasoning with the truth predicate is integrated into our belief system.

Reflection principles express our trust in theories. In this vein, Halbach (2001, p. 1963) states that our *trust* in *TB* is expressed by a uniform reflection principle for *TB*. We can (but do not need to) reflect on our trust in basic arithmetic or in a basic truth theory. This would result in an explicit acceptance of a reflection principle. The explicit theory that is trusted would thereby be enlarged. The trust in this extended theory is implicit in the acceptance of the extended theory. This trust can (but need not be) be articulated in a second reflective moment.

Compare this with reliance on *memory*. At some level of abstraction, memory can be seen as a box. When we rely on our memory, we take things from this box and integrate them into our system of occurrent beliefs. This is what it means to rely on memory. But *that* memory is reliable is not to be found in the memory-box. It could not be![17] Similarly, our basic theory of arithmetic can be seen as a box. When we do arithmetic (calculate and prove propositions), we take things from the box. We integrate these things into our belief system, and thus rely on them. But *that* basic arithmetic is reliable is not, and cannot be, found in the box.

Implicit commitment to reflection principles is implicit in our explicit and unqualified *acceptance* of a theory *S*. Unqualified acceptance of *S* is close to what Franzén (2004, p. 207) calls *acceptance of S as sound*. Franzén glosses acceptance of *S* as sound as 'accepting *that* the axioms of *S* are true' (2004, p. 213; our emphasis). But this phraseology suggests an explicit propositional attitude (acceptance) towards the truth of *S*, and this is more than what unqualified acceptance of *S* entails. Indeed, accepting *S* in an unqualified way is not the same as accepting a reflection principle for *S*. After all, someone might accept *S* without possessing the concept of truth (for *S*) at all.

Not everyone accepts basic arithmetic, classical logic and a basic set of truth-biconditionals in an unqualified way. Someone who is a constructivist, for instance, will not accept classical logic. Such a person can still *use* classical logic *instrumentally*. For instance, given the constructive provability of the Π_2-conservativeness of classical Peano

[17] For an extensive account of our epistemic warrant for relying on our memory, see Burge (1993, 1998).

 © Horsten and Leigh 2016

arithmetic over Heyting arithmetic, the intuitionist may instrumentally use classical Peano arithmetic to derive intuitionistic arithmetical theorems of restricted complexity. But it will be manifested by the constructivist's assertive practice that she does not fully accept classical logic. Therefore she is not entitled to belief in the uniform reflection principle for classical basic arithmetic. In a related vein, it is possible to accept a theory S in a Hilbertian formalist spirit. This is what Franzén (2004, §14.4) calls *accepting S as consistent*. Someone who accepts S in this sense is implicitly entitled to the consistency of S, but not even to Rfn_S.

More generally, being *justified* in accepting a theory S is a necessary condition for being entitled to accept Rfn_S. Failing to be justified in accepting S can of course come about in multiple ways: S may be false, or one may lack sufficient justification for one of S's axioms… Being justified in accepting S is not quite enough to be entitled to accept Rfn_S, but not much more is needed. One additional thing that *is* needed is to understand that the canonical provability predicate $Prov_S$ expresses provability in S.

So, as is the case with reflection principles for sense experience, memory and logical reasoning, we may be in a situation where we think that we are entitled to accept a proof-theoretic reflection principle (or certain instances of it) when we are not. But, as before, this does not mean that our acceptance of proof-theoretic reflection principles can never be more than qualified acceptance.

Is there, as with sense perception, a possibility of brute error when we accept a proof-theoretic reflection principle for a mathematical theory? This question is difficult to answer. Some argue that our justification for our mathematical theories can be reduced to our understanding of the content of mathematical notions (such as the notion of natural number). If this is so—but this is a big 'if'—then brute error is presumably not possible.[18] If, on the other hand, the subject matters of mathematical theories are independently existing platonic realms of which we have fallible mathematical intuition which is somehow akin to sense perception, then brute error is at least to some extent possible. But this is an equally big 'if'. In sum, whether brute error is possible in relying on mathematical theories that we accept, and in explicitly accepting reflection principles for them, depends on larger issues in the

[18] Burge thinks that *simple* mathematical truths are immune to brute error in this way: see Burge (2007, p. 198).

 © Horsten and Leigh 2016

philosophy of mathematics that are of immense importance, but to which we have nothing new to contribute in this article.

It is *possible* to justify uniform reflection for a theory such as *TB*. If one believes in Zermelo-Fraenkel set theory, for instance, and the basic satisfaction-biconditionals for this theory, then one can prove (and thus justify) the soundness of *TB*. The point is just that one *can* legitimately accept uniform reflection for *TB* without having a justification for it. This view seems in line with Feferman's view on the question of the epistemic warrant for reflection principles:

> The idea of an autonomous progression more nearly approximates the process of finding out what is implicit in accepting a basic system L_1, i.e., of what one ought to accept, *on the same fundamental grounds*, when one accepts L_1. (Feferman 1988, p. 131; our emphasis)

There are additional reasons why our warrant for reflection principles cannot be of any of the typical kinds. Reflection principles cannot be *deduced* from the theory upon which the reflection principle reflects. But it is also not possible to take reflection principles to be *axioms* in the sense that, for instance, the power set axiom is an axiom.

According to many, the power set axiom is justified by appeal to a form of *intuition*. We vaguely 'see' a model (the cumulative hierarchy of sets, generated in stages) and realize that the power set axiom is true in it. But the occurrence of the word 'true' in the previous sentence indicates that we do not want to say something similar for reflection principles. For it would mean motivating reflection principles by appeal to the notion of truth. Our position would thereby collapse into the Ketland-Shapiro position, and then the game would be lost. In the type-free setting, it seems clear that the Ketland-Shapiro line is not very plausible. Presumably the uniform reflection principle for our most encompassing theory (including our most encompassing truth theory) holds. So, according to the Ketland-Shapiro line, it should be *justified* on the basis of truth principles, and this justification should presumably take roughly the same form as the canonical justification the reflection principles of sub-theories (such as *TB*) of our overall theory. But for all too familiar Gödelian reasons, we have no truth principles from which we can justify the reflection principle for our most encompassing theory.

According to others, the power set axiom is 'analytic' in some full-blown, Gödel-like sense. But this does not seem to work for reflection principles. The required conceptual connection between provability *in a theory* and truth is just not available. The corresponding conceptual

 © Horsten and Leigh 2016

connection between *informal* provability (not tied to a specific system) and truth *is* there, but that is an entirely different matter.

In any case, reflection principles are *too specific* (tied to a particular theory) to be candidates for being axioms in the true sense of the word, that is, basic principles.[19]

Nonetheless, there may be reason to think that *uniform* reflection principles derive from more basic principles. Kreisel once argued that we believe in the first-order induction scheme (for the language of arithmetic) *because* we believe in the second-order induction axiom (Kreisel 1967). A related train of thought would not go as far as this, but instead, say, argue that we believe in the first-order induction scheme because we believe in the *axiom*

$$\forall \phi \in \mathcal{L}_{PA} : (T\phi(0) \wedge T(\forall y.\ \phi(y) \rightarrow \phi(y+1))) \rightarrow \forall x T\phi(x)$$

If this is plausible, then one might also say that we believe in all instances of the scheme RFN_S *because* we believe in the global reflection *axiom* GRP_S. Thus—or so the argument goes—global reflection *justifies* uniform reflection.

It is hard to evaluate this line of reasoning: what is the evidence for the 'because'? But even if it is found cogent, then the substance of what we have argued for in this article stands. What we have claimed for uniform reflection will then apply instead to global reflection. It will remain the case that the compositionality of truth is contained in disquotational principles for truth, modulo reflection principles; they will then just be global instead of uniform reflection principles.

7. Other roads from disquotation to compositionality

It is fair to say that the doctrine that disquotationalism cannot explain the compositional nature of truth counts as the *received* view. Nevertheless, we are not the only ones to dissent from it. There are some authors who have defended the view that there is a justificatory road from disquotational truth axioms to axioms that express the compositionality of truth. We will discuss three such views here—Field (2006), Halbach (2001) and Halbach (2002)—and explain how the viewpoint that we advocate differs from theirs.

Hartry Field (2006) has shown how compositional axioms can be derived from a specific schematic way of expressing the disquotational

[19] In a similar vein, Donald Martin (1998, p. 227) argues that determinacy principles are too specific to ever be candidates for being basic axioms.

 © Horsten and Leigh 2016

viewpoint. In a nutshell, and simplifying greatly, his theory goes as follows. In Field's account, Tarski-biconditionals are expressed in the language of truth as $Tp \leftrightarrow p$, where p is a *schematic* letter. Then, aside from the usual rules for reasoning in languages with schematic letters, Field introduces a new rule for reasoning with schematic letters, which (roughly) says that if $\Phi(p)$ has been derived, then we are allowed to conclude $\forall \phi \in \mathcal{L} : \Phi(\phi)$, for some language $\mathcal{L}$. So, in particular, since in a schematic disquotational theory we can derive $\neg Tp \leftrightarrow T\neg p$, this rule of inference allows us to derive the compositionality of negation.

The admissibility of this new inference rule indicates that, in effect, a scheme of the form $Tp \leftrightarrow p$ can be read as a *sentence*,

$$\Pi p : Tp \leftrightarrow p,$$

which substitutionally quantifies over propositions of a certain kind. But it is fairly commonly held that substitutional quantification must be explained in terms of a compositional notion of truth. So Field's proposal is particularly vulnerable to Halbach's critique that 'a language embracing substitutional quantification can be considered as a notational variant of the language with the truth predicate satisfying the Tarskian clauses for truth' (Halbach 2002, §5).

In Halbach (2001), an account is given which bears some resemblance to the theory that we are proposing in the present article. Halbach starts by defining the notion of *truth-analyticity*: 'A is truth-analytic iff A is logically implied by the disquotation sentences' (Halbach 2001, p. 1962). Then he argues that RFN_{TB_0} expresses what is *analytical in the concept of truth*. Moreover, he claims that RFN_{TB_0}, rather than TB_0, should be taken to be the disquotationalist's truth theory:

> The formalization of the disquotationalist standpoint by a reflection scheme takes into account that the disquotationalist does not only claim the disquotation sentences, but that he also claims something about them, namely, that they govern the meaning of the truth predicate. (Halbach 2001, p. 1963)

And, as mentioned before, Halbach observes that within the context of Peano arithmetic, RFN_{TB} entails the typed compositional axioms for truth.

What are we to make of this? First of all, we observe that $PA + \mathrm{RFN}_{TB_0}$ is not a collection of truth-biconditionals. So it is hard to see how this is an expression of (rather than a redefinition of) disquotationalism. Second, it is not clear that RFN_{TB_0} can be seen

 © Horsten and Leigh 2016

as a good formalization of the notion of truth-analyticity. Halbach writes:

> In general, I do not propose that axioms of a theory should be replaced by the uniform reflection principles for that theory. For the transition from a theory to a reflection principle for that theory requires an argument. In the case of disquotationalism this is provided by paying attention to the modal status of the disquotation sentences, i.e., by arguments for their analyticity. (Halbach 2001, p. 1963)

We argued in the previous section that if one is fully committed to a theory, then the transition from the theory to the reflection principle for it does not require an *argument*. But, aside from this, even if Halbach is right that truth-analyticity is the *justification* for RFN_{TB}, it does not follow that (and it does not seem correct that) RFN_{TB} is an acceptable *formalization* of truth-analyticity.

Halbach implicitly seems to take this point in his article 'Modalized Disquotationalism', which appeared somewhat later (Halbach 2002). In this article, the compositional axioms are derived from a theory that contains axioms and rules governing a *necessity predicate*, plus a disquotational scheme stating that the Tarski-biconditionals for the language of arithmetic are necessary. So this proposal can be seen as an alternative articulation of the idea that motivates the Halbach (2001) view. After all, nothing precludes interpreting the notion of necessity in Halbach (2002) as 'analytic in the Tarski-biconditionals'.[20] Nevertheless, this theory is also less than satisfactory. First, on the account under investigation it still cannot be said that the compositionality of truth is contained in disquotational axioms. Instead, it is a combined account of necessity and of truth that yields the compositionality of truth. Second, the proposed theory contains seemingly *ad hoc* restrictions on the principles of necessity. In particular, the proposed theory denies the principle that if a sentence necessarily holds, then it is true (Halbach, 2002, §2).[21]

[20] Perhaps this interpretation is suggested in an earlier article by Halbach. In Halbach (2000), he writes, '[O]ne might have the notion of provability from the disquotation sentences, that is, the notion of truth-analyticity, available without believing the soundness of the consequences. The distinctive feature of somebody who believes in the disquotation principle—like the disquotationalist—is his belief in the *soundness* of what I call truth-analyticity' (p. 169). So, 'Disquotationalism is *not* simply an axiomatization of truth but a theory about an axiomatic system of truth ... [D]isquotationalism is not only a theory of truth but a theory of the modal status of the disquotation sentences as well' (p. 170).

[21] Of course, in the light of the Kaplan-Montague paradox for necessity predicates (Kaplan & Montague 1960), either the reflexivity axiom $\Box\phi \rightarrow \phi$ or the necessitation rule for $\Box$ has to be restricted in order to avoid contradiction.

 © Horsten and Leigh 2016

8. Open questions

The conclusion that we have arrived at is that the concept of truth is simple because at bottom disquotational, but reflection is complicated. It has long been known that from a proof-theoretic point of view, reflection hierarchies are complicated and interesting.[22] But we have argued that the process of reflection is also *conceptually* complicated, and that it is intriguing from an epistemological point of view.

In this article we have explored the result of one or two reflective moments applied to a natural typed or type-free collection of truth-biconditionals. But, as we have mentioned, the full implicit content of a theory is not thereby revealed. Instead, it is given by the reflexive closure of the theory in the sense of Feferman. Thus it would seem important to have an informative description of the reflexive closure of *TB* and of *PTB*. Moreover, it is an interesting question whether the Friedman-Sheard theory (as first formulated in Friedman and Sheard 1987) can be obtained by reflection from a natural collection of Tarski-biconditionals.

9. Technical appendix

In this section we provide proofs of the main results outlined above. We begin by fixing notations and definitions referred to earlier.

Any theory of truth requires some background theory of syntax in which the basic syntactic operations corresponding to the manipulation of formulae can be formalized. This requires only a modicum of arithmetic, enough to define a Gödel coding of the prescribed language and simple operations on Gödel codes corresponding to the sentence-building operations and substitution. For this purpose we utilize the theory *EA* of *elementary arithmetic*:[23]

Definition 1. *EA is the theory of arithmetic comprising the basic axioms of PA, including exponentiation, and induction for* Δ_0*-formulae.*

[22] The *locus classicus* for this subject is Feferman (1962).

[23] It is worth noting that *EA* does not represent the *minimal* subsystem of arithmetic that can be used as a base theory for the results of the present paper. This accolade would likely go to Buss's theory of bounded arithmetic S^1_2, whose provably$_1$ total functions are the polytime computable functions. Via an 'efficient' coding of $\mathcal{L}^+$ (see, for example, Hájek and Pudlák 1998, ch. 5), the relevant work, in particular the version of theorem 8 below, can be established with S^1_2 in place of *EA*. Moreover, it is known that S^1_2 permits a finite axiomatization (Ferreira and Ferreira 2013).

 © Horsten and Leigh 2016

As with *PA*, we assume *EA* is formulated in the extended language $\mathcal{L}^+ = \mathcal{L}_T \cup \mathcal{L}_P$ comprising $\mathcal{L}_T$ as well as unary predicates T and F. Nevertheless the induction schema of both theories is restricted to the *truth-free* language (denoted $\mathcal{L}_{PA}$) in which neither predicate T nor F may occur. Thus for each theory it is only the rules and axioms of classical logic that apply to the extended language.

EA, although weak, is sufficiently strong to allow the formulation of the syntactic matters that are required for formalizing standard meta-theoretic arguments of formal theories (sequences, Gödel coding, substitution, etc.). In particular, we can fix a Gödel coding $e \mapsto \ulcorner e \urcorner$ of expressions in the language $\mathcal{L}^+$ such that the functions

$$\dot{\wedge} : (\ulcorner \phi \urcorner, \ulcorner \psi \urcorner) \mapsto \ulcorner \phi \wedge \psi \urcorner \qquad \dot{\forall} : (\ulcorner v \urcorner, \ulcorner \phi \urcorner) \mapsto \ulcorner \forall v \phi \urcorner$$
$$\dot{\vee} : (\ulcorner \phi \urcorner, \ulcorner \psi \urcorner) \mapsto \ulcorner \phi \vee \psi \urcorner \qquad \dot{\exists} : (\ulcorner v \urcorner, \ulcorner \phi \urcorner) \mapsto \ulcorner \exists v \phi \urcorner$$
$$\dot{\neg} : \ulcorner \phi \urcorner \mapsto \ulcorner \neg \phi \urcorner \qquad \dot{=} : (\ulcorner s \urcorner, \ulcorner t \urcorner) \mapsto \ulcorner s = t \urcorner$$
$$sub : (\ulcorner \phi \urcorner, \ulcorner v \urcorner, \ulcorner s \urcorner) \mapsto \ulcorner \phi(s/v) \urcorner$$

are provably total in *EA*.[24] Let $\bar{x}$ denote the xth numeral: $\bar{0} = 0$ and $\overline{x+1} = s(\bar{x})$. For particular applications of the substitution function, it is appropriate to introduce abbreviations: $\overset{T}{\cdot}(x)$ and $\overset{F}{\cdot}(x)$ denote the terms $sub(\ulcorner T(v) \urcorner, \ulcorner v \urcorner, x)$ and $sub(\ulcorner F(v) \urcorner, \ulcorner v \urcorner, x)$ respectively; $subn(\ulcorner \phi \urcorner, \ulcorner v \urcorner, x) = \ulcorner \phi(\bar{x}/v) \urcorner$; $\ulcorner \phi(\dot{v}) \urcorner = subn(\ulcorner \phi \urcorner, \ulcorner v, v \urcorner)$; and for $n > 1$,

$$\ulcorner \phi(\dot{v}_1, \ldots, \dot{v}_n) \urcorner := subn(\ulcorner \phi(\dot{v}_1, \ldots, \dot{v}_{n-1}) \urcorner, \ulcorner v_n \urcorner, v_n)$$

With the above operations, it is straightforward to define predicates $Form_{\mathcal{L}}(x)$, $Sent_{\mathcal{L}}(x)$ and $Term_{\mathcal{L}}(x)$ that represent (in *EA*) the property of being a code of a formula, sentence and term of $\mathcal{L}$. Finally, we also require a valuation function $^\circ : \ulcorner s \urcorner \mapsto s$ that, given the code of a term, returns the value of the term. This function is not provably total in *EA*, but is so in *PA*, which is where we make use of it.

With our formalization of syntax now fixed, we can precisely formulate the uniform reflection principle for a theory S and iterations thereof.

Definition 2. *Let S be a theory in the language $\mathcal{L}^+$ with an elementary decidable set of axioms, and let $Prov_S(x)$ be a standard provability predicate for S formalized within EA.* RFN_S *is the collection of*

[24] We adopt the usual convention of identifying the closed term $\ulcorner e \urcorner$ with its value in the standard model. This should not be confused with the valuation function $^\circ$ introduced below, which, given a number m, determines the term s such that $m = \ulcorner s \urcorner$ and returns the value of s.

 © Horsten and Leigh 2016

formulae

$$\forall x_1 \ldots \forall x_n(Prov_S(\ulcorner \phi(\dot{x}_1, \ldots, \dot{x}_n)\urcorner) \rightarrow \phi)$$

for ϕ, *a formula of* $\mathcal{L}^+$ *with free variables among* $x_1, \ldots, x_n$. Let $S_1 = EA + \mathrm{RFN}_S$ and $S_{n+1} = EA + \mathrm{RFN}_{S_n}$.

The following result, originally due to Kreisel and Levy (1968), neatly demonstrates the connection between induction and reflection.[25]

Theorem 8. $EA_1 = PA + \mathrm{Ind}(\mathcal{L}^+)$

As the results in this paper make extensive use of the techniques used to prove theorem 8, we outline the proof.

Proof. That PA is a sub-theory of EA_1 requires proving that each instance of $\mathcal{L}^+$-induction is derivable from an instance of the reflection principle. Let $\phi(x)$ be a formula of $\mathcal{L}^+$ and let $\psi(x) = \phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x + 1)) \rightarrow \phi(x)$. Within EA, it is possible to prove that for every n, a derivation in EA of $\psi(\bar{n})$ can be transformed into a derivation of $\psi(\overline{n+1})$. By Δ_0-induction, we deduce $EA \vdash \forall x Prov_S(\ulcorner \psi(\dot{x})\urcorner)$, and so $EA_1 \vdash \forall x \psi$. As $\forall x \psi$ is provably (in EA) equivalent to the axiom of induction for ϕ, we are done.

The converse argument requires more delicate proof-theoretic analysis. EA may not be finitely axiomatizable, but $I\Sigma_1 = EA + \mathrm{Ind}(\Sigma^0_1)$, the extension of EA by induction for Σ^0_1 formulae (from $\mathcal{L}_{PA}$), can be axiomatized as a single Σ^0_3 formula by coding the induction schema as a single instance using a Σ^0_1 partial truth predicate (see, for example, Hájek and Pudlák 1998, theorem 2.52). Hence, for every theorem $\phi \in \Sigma_n$ of $I\Sigma_1$ (say $n \geq 3$) there is a derivation of ϕ in a sequent calculus formulation of $I\Sigma_1$ with a cut rule, in which axioms involve only Σ^0_3 formulae. Standard cut elimination for first-order logic yields a derivation of ϕ with cuts-only axioms, and so the sub-formula property implies that the whole derivation consists solely of Σ^0_n formulae. This partial cut elimination argument is formalizable within $I\Sigma_1$. Within PA therefore, a partial truth predicate $Tr^*(x)$ can be defined such that $PA \vdash Tr^*(\ulcorner \phi(\dot{x})\urcorner) \leftrightarrow \phi(x)$ for each Σ^0_n formula ϕ, and

$$PA + \mathrm{Ind}(\mathcal{L}^+) \vdash \forall x(Sent_{\Sigma_n}(x) \wedge Prov_{EA}(x) \rightarrow Tr^*(x))$$

we deduce $PA + \mathrm{Ind}(\mathcal{L}^+) \vdash \mathrm{RFN}_{EA}$. □

[25] For a thorough discussion of this result and its philosophical significance, see Dean (2015).

 © Horsten and Leigh 2016

9.1 Typed truth

Definition 3. *The theories TB_0 and UTB_0 are the extensions of EA by, respectively, the local and uniform truth-biconditionals for $\mathcal{L}_{PA}$. That is, TB_0 contains the axiom $\phi \leftrightarrow T(\ulcorner \phi \urcorner)$ for each closed $\mathcal{L}_{PA}$-formula ϕ, and UTB_0 contains the axiom $\phi(x_1, \ldots, x_n) \leftrightarrow T(\ulcorner \phi(\dot{x}_1, \ldots, \dot{x}_n) \urcorner)$ for each $n \geq 0$ and each formula ϕ with at most $x_1, \ldots, x_n$ free. By TB and UTB we denote the respective theory with induction extended to all formulae of the language.*[26]

Theorem 9. *UTB is a sub-theory of TB_1.*

Proof. For each closed formula $\forall x \phi(x)$ of $\mathcal{L}_{PA}$ we have, provably in *EA*,

$$TB_0 \vdash T(\ulcorner \phi(\bar{n}) \urcorner) \leftrightarrow \phi(\bar{n})$$

for every n, whence reflection implies $TB_1 \vdash \forall x(T\ulcorner \phi(\dot{x}) \urcorner \leftrightarrow \phi(x))$. □

Definition 4. *CT is the $\mathcal{L}_T$ theory extending PA by induction for all $\mathcal{L}_T$ formulae and the following axioms:*

(1) $\forall s_1, s_2[Term_{\mathcal{L}}(s_1) \wedge Term_{\mathcal{L}}(s_2) \rightarrow (T(s_1 \doteq s_2) \leftrightarrow \overset{\circ}{s_1} = \overset{\circ}{s_2})]$

(2) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_T}(\alpha_1 \dot{\wedge} \alpha_2) \rightarrow (T(\alpha_1 \dot{\wedge} \alpha_2) \leftrightarrow T\alpha_1 \wedge T\alpha_2)]$

(3) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_T}(\alpha_1 \dot{\wedge} \alpha_2) \rightarrow (T(\alpha_1 \dot{\vee} \alpha_2) \leftrightarrow T\alpha_1 \vee T\alpha_2)]$

(4) $\forall \alpha[Sent_{\mathcal{L}_T}(\alpha) \rightarrow (T(\dot{\neg} \alpha) \leftrightarrow \neg T\alpha)]$

(5) $\forall \alpha[Sent_{\mathcal{L}_T}(sub_x(\alpha, \bar{0})) \rightarrow (T(\dot{\forall} x \alpha) \leftrightarrow \forall x\, T(sub_x(\alpha, x))]$

(6) $\forall \alpha[Sent_{\mathcal{L}_T}(sub_x(\alpha, \bar{0})) \rightarrow (T(\dot{\exists} x \alpha) \leftrightarrow \exists x\, T(sub_x(\alpha, x)))]$

Halbach (2001, lemma 4.2) established:

Theorem 10. *CT and UTB_1 are identical theories.*[27]

Proof. We outline Halbach's argument, as it demonstrates how compositional axioms arise through reflection.

Arguing informally within TB_1, we observe that for all $\mathcal{L}$-sentences ϕ and ψ,

$$TB_0 \vdash T(\ulcorner \phi \wedge \psi \urcorner) \leftrightarrow T(\ulcorner \phi \urcorner) \wedge T(\ulcorner \psi \urcorner)$$

[26] In general, we append a subscript 'o' to theories to emphasize that induction is *not* extended; truth theories without subscripts will feature induction for the whole language.

[27] The theories *CT* and UTB_1 are denoted by $PA(S)$ and AT respectively in Halbach (2001).

 © Horsten and Leigh 2016

An application of reflection therefore yields the compositional axiom for conjunction. A similar argument implies the compositional axiom for disjunction and atomic predicates. The quantifier axioms follow by an analogous argument though the theory UTB_1 is needed in place of TB_1. Thus all axioms of CT are derivable in UTB_1.

To deduce that the axioms of UTB_1 are derivable in CT, one uses a slight generalization of Kreisel and Lévy's argument as described in Halbach (2001). It is clear that UTB_0 is a sub-theory of CT. Crucially, there is a finite set of axioms of CT that suffice to derive the axioms of UTB_0, namely, the extension of $I\Sigma_1$ by the six axioms in definition 4. Utilizing partial cut elimination for this theory as well as extensions to the partial truth predicates that are available in CT, it follows that the schema RFN_{UTB_0} is derivable in CT. □

Combining this result with theorem 9 we conclude:

Theorem 11. *CT is a proper sub-theory of TB_2.*

Proof. That CT is a sub-theory of TB_2 follows from the previous two theorems. To conclude that the reflection principle for TB_1 is not derivable in CT we resort to known facts from the proof-theoretic analysis of CT.

Since induction for the whole language is already derivable in TB_1, it follows that within TB_2 transfinite induction through the ordinal ε_0 is derivable for every $\mathcal{L}_T$ formulae. It is well known (see, for example, Feferman 1991) that within CT, transfinite induction for $\mathcal{L}_T$ formulae is derivable only for ordinals strictly below ε_0.[28] Thus CT must be a proper sub-theory of TB_2. □

9.2 Type-free truth

Recall the language $\mathcal{L}_P$ that contains all terms and relations of $\mathcal{L}_T$, and

(1) for each atomic predicate A of $\mathcal{L}_T$, a fresh predicate symbol $\bar{A}$ in $\mathcal{L}_P$ of the same arity as A; each of A and $\bar{A}$ is referred to as the *dual* of the other,

(2) the connectives $\wedge$ and $\vee = \bar{\wedge}$, and quantifiers $\forall$ and $\exists = \bar{\forall}$.

Similarly we call $\wedge$ and $\vee$ dual as well as $\forall$ and $\exists$.

We denote the dual of T by F. In this way $\mathcal{L}_P$ can be considered as the negation-free sub-language of $\mathcal{L}_{T,F}$ (the extension of $\mathcal{L}$ by two unary

[28] In contrast to PA_T, however, CT does prove transfinite induction to ordinals below ε_0 for formulae of $\mathcal{L}_{PA}$.

 © Horsten and Leigh 2016

predicate symbols, T and F) in which F is identified with $\neg T$. Each formula ϕ of $\mathcal{L}_P$ has a dual $\bar{\phi}$ that results from recursively replacing each connective, quantifier and predicate symbol in ϕ by its dual.

Definition 5. *TFB_0 is the $\mathcal{L}_{T,F}$-theory extending EA by the truth and falsity biconditionals for all formulae in $\mathcal{L}_P$: the collection of axioms*

$$T(\ulcorner \phi \urcorner) \leftrightarrow \phi F(\ulcorner \bar{\phi} \urcorner) \leftrightarrow \phi$$

for sentences ϕ of $\mathcal{L}_P$.
$UTFB_0$ extends TFB_0 by the equivalences

$$T(\ulcorner \phi(\dot{v}_1, \ldots, \dot{v}_n) \urcorner) \leftrightarrow \phi F(\ulcorner \overline{\phi(\dot{v}_1, \ldots, \dot{v}_n)} \urcorner) \leftrightarrow \phi$$

for each formula ϕ, with at most $v_1, \ldots, v_n$ free. KF is the $\mathcal{L}_{T,F}$-theory extending PA by induction for all $\mathcal{L}_{T,F}$ formulae and the following twelve axioms.

(1) $\forall s_1, s_2[Term_{\mathcal{L}}(s_1) \wedge Term_{\mathcal{L}}(s_2) \rightarrow (T(s_1 \doteq s_2) \leftrightarrow s_1^\circ = s_2^\circ)]$

(2) $\forall s_1, s_2[Term_{\mathcal{L}}(s_1) \wedge Term_{\mathcal{L}}(s_2) \rightarrow (F(s_1 \doteq s_2) \leftrightarrow s_1^\circ \neq s_2^\circ)]$

(3) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (T(\alpha_1 \wedge \alpha_2) \leftrightarrow T\alpha_1 \wedge T\alpha_2)]$

(4) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (T(\alpha_1 \vee \alpha_2) \leftrightarrow T\alpha_1 \vee T\alpha_2)]$

(5) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (F(\alpha_1 \vee \alpha_2) \leftrightarrow F\alpha_1 \wedge F\alpha_2)]$

(6) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (F(\alpha_1 \wedge \alpha_2) \leftrightarrow F\alpha_1 \vee F\alpha_2)]$

(7) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (T(\forall x\alpha) \leftrightarrow \forall x\, T(sub_x(\alpha, x)))]$

(8) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (T(\exists x\alpha) \leftrightarrow \exists x\, T(sub_x(\alpha, x)))]$

(9) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (F(\forall x\alpha) \leftrightarrow \exists x\, F(sub_x(\alpha, x)))]$

(10) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (F(\exists x\alpha) \leftrightarrow \forall x\, F(sub_x(\alpha, x)))]$

(11) $\forall s[Term_{\mathcal{L}}(s) \rightarrow (T(T\, s) \leftrightarrow T(s^\circ)) \wedge (T(F\, s) \leftrightarrow F(s^\circ))$

(12) $\forall s[Term_{\mathcal{L}}(s) \rightarrow (F(T\, s) \leftrightarrow F(s^\circ)) \wedge (F(F\, s) \leftrightarrow T(s^\circ))$

The reader may be worried by the fact that the sentential quantifiers in axioms 3–10 of *KF* are restricted to the language $\mathcal{L}_P$, which does not contain the negation symbol, and not to either the languages $\mathcal{L}_{T,F}$ or $\mathcal{L}_T$, as is most common in the literature. The reason is that it is only when viewed as a theory of truth for the language $\mathcal{L}_P$ that the axioms of *KF* can be described as compositional with respect to *all* the

 © Horsten and Leigh 2016

connectives—*KF* (in any of its incarnations) is inconsistent with the compositional axiom for negation, $Sent_{\mathcal{L}}(\alpha) \rightarrow T(\dot{\neg}\alpha) \leftrightarrow \neg T(\alpha)$ (where $\mathcal{L}$ is any one of $\mathcal{L}_T$, $\mathcal{L}_{T,F}$ or $\mathcal{L}_P$). In the languages $\mathcal{L}_T$ and $\mathcal{L}_{T,F}$, occurrences of the negation symbol under the truth predicate are afforded a different interpretation from those outside, being either unspecificied (as in Halbach 2001, for instance) or (as in Feferman 1991) interpreted as behaving as our duality operator, exchanging the roles of 'true' and 'false', etc.

We begin by noting that, taken in isolation, the truth- and falsity-biconditionals are model-theoretically weak.

Theorem 12. *$UTFB_0$, and hence also TFB_0, is semantically conservative over EA.*

Proof. We present here the proof that every model of *EA* can be extended to a model of TFB_0 via a variant of Kripke's fixed-point construction; a proof of the same result for $UTFB_0$ follows the same lines and is left to the reader.

For an $\mathcal{L}$-structure $\mathfrak{M}$, R, S subsets of the domain of $\mathfrak{M}$ (i.e. $R, S \subseteq |\mathfrak{M}|$), and χ a formula of $\mathcal{L}_{T,F}$, let $\langle \mathfrak{M}, R, S\rangle \models \chi$ be defined according to the usual Tarskian satisfaction rules for classical logic with R interpreting the extension of the truth predicate T and S interpreting F. Notice that for χ in $\mathcal{L}_P$, $R_0 \subseteq S_0 \subseteq |\mathfrak{M}|$ and $R_1 \subseteq S_1 \subseteq |\mathfrak{M}|$ we have

$$\langle \mathfrak{M}, R_0, R_1\rangle \models \chi \text{ implies } \langle \mathfrak{M}, S_0, S_1\rangle \models \chi$$

Let $\bar{R} = \{\ulcorner \bar{\chi} \urcorner | \chi \in R\}$. We now define a sequence of sets R_α indexed by countable ordinals:

$$R_0 = \varnothing$$
$$R_{\alpha+1} = \{\ulcorner \psi \urcorner | \psi \text{ is a sentence of } \mathcal{L}_P \text{ and } \langle \mathfrak{M}, R_\alpha, \bar{R}_\alpha\rangle \models \psi\}$$
$$R_\lambda = \bigcup_{\alpha < \lambda} R_\alpha \text{ (for limit } \lambda)$$

Since $R_\alpha \subseteq R_\beta$ for every $\alpha \leq \beta$, by cardinality considerations there exists κ for which $R_\kappa = R_{\kappa+1}$. Then for a sentence ψ of $\mathcal{L}_P$,

$$\begin{aligned}\langle \mathfrak{M}, R_\kappa, \bar{R}_\kappa\rangle \models \psi &\Leftrightarrow \psi \in R_{\kappa+1} \\ &\Leftrightarrow \langle \mathfrak{M}, R_{\kappa+1}, \bar{R}_{\kappa+1}\rangle \models T(\ulcorner \dot{\psi} \urcorner) \\ &\Leftrightarrow \langle \mathfrak{M}, R_\kappa, \bar{R}_\kappa\rangle \models T(\ulcorner \psi \urcorner)\end{aligned}$$

 © Horsten and Leigh 2016

and similarly for the falsity predicate. Therefore $\langle \mathfrak{M}, R_K, \bar{R}_K \rangle \models TFB_0$.

Cieśliński establishes that induction for the extended language does not increase the proof-theoretic strength of the local positive truth-biconditionals, that is, *PTB* conservatively extends *PA* (Cieśliński 2011). The proof naturally extends to the case of a falsity predicate:

Theorem 13. *$TFB = TFB_0 + \mathrm{Ind}(\mathcal{L}_{T,F})$, and hence also PTB, conservatively extends PA.*

We present a simplified version of Cieśliński's proof below.[29] Before we proceed, observe that the proof of the theorem 12 does not suffice for this result, as the constructed structure $\langle \mathfrak{M}, R_K, \bar{R}_K \rangle$ need not satisfy the extended induction schema $\mathrm{Ind}(\mathcal{L}_{T,F})$. Instead we prove that for each finite set S of $\mathcal{L}_P$ sentences there exists an $\mathcal{L}_{PA}$-definable subset R of $|\mathfrak{M}|$ *and* $\langle \mathfrak{M}, R, \bar{R} \rangle$ satisfies the truth and falsity biconditionals for all formulae in S. The first requirement ensures that the structure $\langle \mathfrak{M}, R, \bar{R} \rangle$ satisfies the induction schema for formulae of $\mathcal{L}_{T,F}$ whenever $\mathfrak{M}$ is a model of $\mathcal{L}_{PA}$-induction.

Proof. Let $\mathfrak{M}$ be an arbitrary model of *PA*. We prove that if ϕ is derivable in $TFB_0 + \mathrm{Ind}(\mathcal{L}_{T,F})$ and ϕ is a formula of $\mathcal{L}_{PA}$, then $\mathfrak{M} \models \phi$.

Suppose ϕ in the language $\mathcal{L}_{T,F}$ is derivable in *TFB*. Let S denote the (finitely many) $\mathcal{L}_P$-formulae such that ϕ is derivable from $EA + \mathrm{Ind}(\mathcal{L}_{T,F})$ extended by the axioms

$$T(\ulcorner \psi \urcorner) \leftrightarrow \psi \qquad\qquad F(\ulcorner \psi \urcorner) \leftrightarrow \psi$$

for each $\psi \in S$. We inductively define

$$R_0 := \varnothing \; R_{n+1} := \{\ulcorner \psi \urcorner \mid \psi \in S \wedge \langle \mathfrak{M}, R_n, \bar{R}_n \rangle \models \psi\}$$

Since S contains only formulae of $\mathcal{L}_P$ we have $R_n \subseteq R_{n+1}$ for every n. As S is finite, there is some k (indeed $k = |S|$ suffices) such that $R_k = R_{k+1}$. Thus we deduce that $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle$ satisfies $\psi \leftrightarrow T(\ulcorner \psi \urcorner)$ and $\psi \leftrightarrow F(\ulcorner \psi \urcorner)$ for each $\psi \in S$, as in the proof of the previous theorem. We claim also $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle \models \mathrm{Ind}(\mathcal{L}_{T,F})$.

Define formulae $T_i(x)$ and $F_i(x)$ as

$$T_0(x) \leftrightarrow \bot \qquad\qquad F_i(x) \leftrightarrow \bot$$

[29] It is worth remarking that unlike the proof presented here, Cieslinski's own argument implies a more general result than stated in theorem 13, namely, that every recursively saturated model of PA can be extended to a model of PTP (and indeed also PTB).

 © Horsten and Leigh 2016

$$T_{i+1}(x) \leftrightarrow \bigvee_{\psi \in S} (x = \ulcorner \psi \urcorner \wedge \psi^i) \quad F_{i+1}(x) \leftrightarrow \bigvee_{\psi \in S} (x = \ulcorner \psi \urcorner \wedge \psi^i)$$

where ψ^i denotes the result of replacing in ψ the predicates T and F by T_i and F_i respectively.

$$R_k = R_{k+1} = \{\ulcorner \chi \urcorner | \chi \in S \wedge \mathfrak{M} \models \chi^k\}$$

and $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle \models \mathrm{Ind}(\mathcal{L}_{T,F})$. Thus $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle \models \phi$ and so $\mathfrak{M} \models \phi^k$. Since $\phi^k = \phi$ if ϕ is in the language of $\mathcal{L}_{PA}$, we are done. □

In contrast to the case of theorem 12, in the above theorem *TFB* cannot be replaced by *UTFB*. Since *PUTB* is interpretable in *UTFB*, it follows that *KF* is also interpretable in *UTFB* (see below). The proof above breaks down in this case, because choosing a finite collection of *uniform* biconditionals does not ensure a fixed point is reached within a finite number of steps. Consider, for example, a formula $\phi(x)$ with only x free (constructed via diagonalization) such that

$$\phi(x) \leftrightarrow x = 0 \vee (x > 0 \wedge T(\ulcorner \phi(\dot{x} - 1) \urcorner))$$

If we fix the case that $\mathfrak{M} = N$ is the standard model of arithmetic, it is natural to choose $S = \{\ulcorner \phi(\bar{n}) \urcorner < \omega\}$ and define the sets $R_0 = \varnothing$, R_1, $R_2, \ldots$ as before. In this case, $\phi(\bar{n}) \in R_{n+1} \backslash R_n$ for every n, so the first structure in the hierarchy satisfying the biconditional $\phi(x) \leftrightarrow T(\ulcorner \phi(\dot{x}) \urcorner)$ will be the limit structure $\langle \mathfrak{M}, \bigcup_{k<\omega} R_k, \bigcup_{k<\omega} \bar{R}_k \rangle$. In general, we cannot expect the fixed point to be definable.

We now turn our attention to extensions of the basic type-free theories of truth by reflection principles.

Theorem 14 (Halbach 2009, theorem 5.1). *KF is directly interpretable in PUTB.*

The interpretation is a trick of diagonalization: one constructs, via the diagonalization lemma, a formula $\phi(x)$ such that

$$\phi(x \dot{\vee} y) \leftrightarrow T(\ulcorner \phi(\dot{x}) \urcorner) \vee T(\ulcorner \phi(\dot{y}) \urcorner)$$
$$\phi(\dot{\forall} vx) \leftrightarrow \forall y\, T(\ulcorner \phi(subn(\dot{x}, \dot{v}, \dot{y})) \urcorner)$$
$$\phi(\dot{T}(x)) \leftrightarrow T(\ulcorner \phi(\dot{x}^\circ) \urcorner)$$
$$\ldots$$

© Horsten and Leigh 2016

Observe that ϕ is positive in T, so one has $PUTB \vdash \phi(x) \leftrightarrow T(\ulcorner \phi(\dot{x}) \urcorner)$, and hence the formula ϕ may serve as the interpretation of the *KF*-truth predicate.

Further reflection on the positive truth-biconditionals does, in a sense, yield the *KF* axioms:

Theorem 15 (Halbach 2001, theorem 5.2). *There are recursive functions f_T and f_F such that the translation that maps $T(s)$ to $T(f_T(s))$ and $F(s)$ to $F(f_F(s))$ is a direct interpretation of KF in $PUTB_1$.*

As a corollary, the same interpretation yields an embedding of *KF* into PTB_2. Since *KF* is formulated in the language $\mathcal{L}_{T,F}$ (both here and in Halbach 2001), it is not a sub-theory of PTB_n for any n. The missing axioms for the falsity predicate can be obtained by starting instead from *TFB*.

Theorem 16. *$KF = UTFB_1$ and KF is a sub-theory of TFB_2.*

Proof. We begin by showing that $KF = UTFB_1$. Both directions follow similar arguments as for theorem 10. As before, all the basic compositional axioms of *KF* are derivable from particular instances of reflection over $UTFB_0$. To deduce the final two axioms we observe that $T(\ulcorner Ts \urcorner) \leftrightarrow T(s)$, and hence,

$$(2) \quad T(\dot{T}(\ulcorner s \urcorner)) \leftrightarrow T(\ulcorner s \urcorner^{\circ})$$

is derivable in *TB* for each term s.[30] But then $UTB_1 \vdash \forall x(Term_{\mathcal{L}}(x) \rightarrow (T(\dot{T}\, x) \leftrightarrow T(x^{\circ})))$ as required, and similarly for the other variations. The proof that $UTFB_1$ is a sub-theory of *KF* is analogous to theorems 8 and 11, and we refer the reader interested in the differences to Halbach (2001, lemma 6.5).[31] As in the typed case, $UTFB_1$ is a sub-theory of TFB_2, whence the second part is established.

The inclusion $KF \subseteq TFB_2$ turns out to be strict:

Theorem 17. *KF is a proper sub-theory of TFB_2.*

[30] Since $^{\circ}$ is not provably total in *EA*, the formula represented in (2) is not, strictly speaking, derivable in *TB*. To be fully precise (and to suffice for the present argument), there exists a formula ϕ such that $PA \vdash \phi(\ulcorner s \urcorner, s)$ for every term s and $PA \vdash \forall x \forall y(Term_{\mathcal{L}}(x) \wedge \phi(x, y) \rightarrow (T(\dot{T}\, x) \leftrightarrow T(y)))$. As $^{\circ}$ is not, strictly speaking, a term in *PA*, this will be the form of *KF* axioms anyway.

[31] Halbach's argument concerns the theory $PUTB_1$ (denoted therein AT^{+}) in place of $UTFB_1$ and an axiomatization of *KF* with only a single truth predicate, but the argument is identical.

 © Horsten and Leigh 2016

Proof. The proof of theorem 14 establishes an $\mathcal{L}$-conservative interpretation of *KF* into the theory *PUTB*, proving that the two theories have the same truth-free consequences. As this latter theory is a sub-theory of TFB_1, it follows that *KF* is $\mathcal{L}$-conservatively interpretable in TFB_1. Since the consistency statement for TFB_1 is derivable in TFB_2, the result follows.[32]

References

Baker, Alan 2005: 'Are There Genuine Mathematical Explanations of Physical Phenomena?' *Mind*, 114, pp. 223–38.

Burge, Tyler 1993: 'Content Preservation'. *Philosophical Review*, 102, pp. 457–88.

——1998: 'Computer Proof, Apriori Knowledge, and Other Minds'. *Philosophical Perspectives, 12: Language, Mind, and Ontology*, pp. 1–37.

——2003: 'Perceptual Entitlement'. *Philosophy and Phenomenological Research*, 67, pp. 503–48.

——2007: 'Self and Self-Understanding'. In Burge 2013, pp. 187–226.

——2013: *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection. Philosophical Essays*, Volume 3. Oxford: Oxford University Press.

Burgess, John P. 1986: 'The Truth is Never Simple'. *Journal of Symbolic Logic*, 51, pp. 663–81.

Cieśliński, Cezary 2010: 'Truth, Conservativeness, and Provability'. *Mind*, 119, pp. 409–22.

——2011: 'T-equivalences for Positive Sentences'. *Review of Symbolic Logic*, 4(2), pp. 319–25.

Davidson, Donald 1967: 'Truth and Meaning'. *In his Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, 1984.

Dean, Walter 2015: 'Arithmetical Reflection and the Provability of Soundness'. *Philosophia Mathematica*, 23(1), pp. 31–64.

Feferman, Solomon 1962: 'Transfinite Recursive Progressions of Axiomatic Theories'. *Journal of Symbolic Logic*, 27(3), pp. 259–316.

——1988: 'Turing in the Land of O(z)'. In Rolf Herken (ed.), *The Universal Turing Machine: A Half-Century Survey*, pp. 113–47. Hamburg: Kammerer & Univerzagt.

[32] The authors wish to thank Volker Halbach, Jeff Ketland, Philip Welch, Marianna Antonutti, Walter Dean and Carlo Nicolai for their valuable comments and suggestions for improvement. The second author was supported by Arts and Humanities Research Council UK grant no. AH/H039791/1.

 © Horsten and Leigh 2016

——1991: 'Reflecting on Incompleteness'. *Journal of Symbolic Logic*, 56, pp. 1–49.

Ferreira, Fernando, and Gilda Ferreira 2013: 'Interpretability in Robinson's Q'. *Bulletin of Symbolic Logic*, 19(3), pp. 289–317.

Field, Hartry 1999: 'Deflating the Conservativeness Argument'. *Journal of Philosophy*, 99, pp. 534–40.

——2006: 'Compositional Principles vs. Schematic Reasoning'. *Monist*, 89, pp. 9–27.

Franzén, Torkel 2004: *Inexhaustibility: A Non-Exhaustive Treatment*. Wellesley, MA: A. K. Peters.

Friedman, Harvey, and Michael Sheard 1987: 'An Axiomatic Approach to Self-Referential Truth'. *Annals of Pure and Applied Logic*, 33, pp. 1–21.

Hájek, Petr, and Pavel Pudlák 1998: *Metamathematics of First-Order Arithmetic*. Berlin: Springer-Verlag.

Halbach, Volker 2000: 'Disquotationalism Fortified'. In André Chapuis and Anil Gupta (eds.), *Circularity, Definition, and Truth*. New Delhi: Munshiram Manohararlal Publishers.

——2001: 'Disquotational Truth and Analyticity'. *Journal of Symbolic Logic*, 66, pp. 1959–73.

——2002: 'Modalized Disquotationalism'. In Volker Halbach and Leon Horsten (eds.), *Principles of Truth*, pp. 75–101. Frankfurt am Main: Dr. Hänsel-Hohenhausen.

——2009: 'Reducing Compositional to Disquotational Truth'. *Review of Symbolic Logic*, 2, pp. 786–98.

——2011: *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

Horsten, Leon 2011: *The Tarskian Turn: Deflationism and Axiomatic Truth*. Cambridge, MA: MIT Press.

Horwich, Paul 1998: *Truth*, 2nd edn. Oxford: Clarendon Press.

Kaplan, David, and Richard Montague, 1960. 'A Paradox Regained'. *Notre Dame Journal of Formal Logic*, 1, pp. 79–90.

Ketland, Jeffrey 1999: 'Deflationism and Tarski's Paradise'. *Mind*, 108, pp. 70–94.

——2005: 'Deflationism and the Gödel Phenomena: Reply to Tennant'. *Mind*, 114, pp. 76–88.

——2010: 'Truth, Conservativeness, and Provability: Reply to Cieśliński'. *Mind*, 119, pp. 423–36.

Kotlarski, H., S. Krajewski ,and A. H. Lachlan 1981: 'Construction of Satisfaction Classes for Nonstandard Models'. *Canadian Mathematical Bulletin*, 24, pp. 283–93.

 © Horsten and Leigh 2016

Kreisel, Georg 1967: 'Informal Rigour and Completeness Proofs'. In Imre Lakatos (ed.), *Problems in the Philosophy of Mathematics*. pp. 138–86. Amsterdam: North-Holland.

——1970: 'Principles of Proof and Ordinals Implicit in Given Concepts'. In A. Kino, J. Myhill, and R. E. Vesley (eds.), *Intuitionism and Proof Theory*. pp. 489–516. Amsterdam: North-Holland.

Kreisel, Georg and A. Lévy 1968: 'Reflection Principles and their Use for Establishing the Complexity of Axiomatic Systems'. *Mathematical Logic Quarterly*, 14(7-12), pp. 97–142.

Lachlan, A. H. 1981: 'Full Satisfaction Classes and Recursive Saturation'. *Canadian Mathematical Bulletin*, 24, pp. 295–7.

Martin, D. A. 1998: 'Mathematical Evidence'. In H. G. Dales and G. Oliveri (eds.), *Truth in Mathematics*, pp. 215–31. Oxford: Oxford University Press.

McGee, Vann 1992: 'Maximal Consistent Sets of Tarski's Schema (T)'. *Journal of Philosophical Logic*, 21, pp. 235–41.

Quine, W. V. 1970: *Philosophy of Logic*. Cambridge, MA: Harvard University Press.

Soames, Scott 2003: *Philosophical Analysis in the Twentieth Century, vol. 2: The Age of Meaning*. Princeton, NJ: Princeton University Press.

Shapiro, Stewart 1998: 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy*, 95, pp. 493–521.

Strollo, Andrea 2013: 'Deflationism and the Invisible Power of Truth'. *Dialectica*, 67, pp. 521–43.

Tarski, Alfred 1935: 'The Concept of Truth in Formalized Languages'. In Tarski 1983, pp. 152–278.

——1983: *Logic, Semantics, Metamathematics: Papers from 1923 to 1938* rev. edn. Translated by J. H. Woodger. Indianapolis: Hackett.

——1986: 'What are Logical Notions?' Edited with an introduction by John Corcoran in *History and Philosophy of Logic*, 7, pp. 143–54.

Tennant, Neil 2002: 'Deflationism and the Gödel Phenomena'. *Mind*, 111, pp. 551–82.

——2005: 'Deflationism and the Gödel Phenomena: Reply to Ketland'. *Mind*, 114, pp. 89–96.

——2010: 'Deflationism and the Gödel Phenomena: Reply to Cieśliński'. *Mind*, 119, pp. 438–50.

Williams, Michael 1988: 'Epistemological Realism and the Basis of Scepticism'. *Mind*, 97, pp. 415–39.

 © Horsten and Leigh 2016

Moral Judgement, Agency and Affect: A Response to Gerrans and Kennett

Chris Zarpentine
Wilkes University
christoph.zarpentine@wilkes.edu

Recently, a number of philosophers and psychologists have drawn on neuroscientific and psychological research on the role of affective processes in moral thinking to provide support for moral sentimentalism. Philip Gerrans and Jeanette Kennett (2010) criticize such 'neurosentimentalist' accounts on the grounds that they focus only on synchronic processes occurring at the time of moral judgement. As a result, these accounts face a dilemma: either they fail to accommodate the connection between moral judgement and agency or they are committed to implausible claims about the moral agency of individuals with damage to the ventromedial prefrontal cortex (vmPFC). I respond to this criticism, arguing that Gerrans and Kennett fail to appreciate the diachronic aspects of affective mechanisms and that they misinterpret the empirical literature on the vmPFC. I argue that neurosentimentalism does have the resources to explain the connection between moral judgement and agency.

Philosophers in the tradition of *rationalism* generally downplay the role of emotion in moral judgement, often claiming that emotions disrupt moral judgement more than they contribute to it. In contrast, *sentimentalism* and various forms of *noncognitivism* and *expressivism* hold that emotions are central to moral judgement. A number of philosophers and psychologists have sought to develop and support sentimentalist views by drawing on a range of neuroscientific and psychological evidence. A recent paper by Philip Gerrans and Jeanette Kennett (2010) (hereafter G&K) aims to criticize such accounts on the grounds that they cannot accommodate the fact that moral judgement is essentially an agential capacity. On their view, these accounts 'leave the moral agent out of the picture' (2010, p. 606).

Here I offer a response to this criticism. After introducing some of the neurosentimentalist views recently put forward and briefly discussing some of the evidence in support of them (§1), I turn to G&K's criticism (§2). They argue that neurosentimentalists views either cannot explain the connection between moral judgement and agency

Mind, Vol. 126 . 501 . January 2017
© Zarpentine 2016
doi:10.1093/mind/fzv202 Advance Access publication 22 October 2016

or imply implausible claims about the agential status of individuals with damage to the ventromedial prefrontal cortex (vmPFC). In the following section (§3), I argue that careful attention to the empirical research undermines G&K's criticism of neurosentimentalism. G&K neglect an important diachronic aspect of affective mechanisms and misinterpret the literature on the function of the vmPFC. When these errors are corrected it is clear that neurosentimentalism has the resources to offer a plausible account both of the connection between moral judgement and agency and of the important role of affect in moral thought and action.

1. Neurosentimentalism

Gerrans and Kennett focus their criticism on accounts that make two kinds of claim. First, they address *neurosentimentalist* accounts of moral judgement: views that draw on neuroscientific work to support the claim that 'moral judgement depends essentially on tacit affective processes' (2010, p. 586). However, they are also interested in criticizing related *rationalist externalist* accounts according to which rational processes may contribute to moral judgement but are, by themselves, motivationally inert. Such accounts are so-called because they deny *internalism*, the view that there is an 'internal' or necessary connection between moral judgement and motivation.

There are several extant accounts that endorse one or both of these two claims. Jonathan Haidt (2001) has developed the *social intuitionist* model of moral judgement, according to which moral judgement generally depends upon quick, automatic, intuitive responses, many of which are affective. In supporting his view, he has drawn on a range of research, particularly in social psychology, that indicates that moral judgement is primarily driven by automatic, affective responses[1] and that consciously controlled reasoning occurs largely as an exercise in post-hoc justification. For example, studies utilizing hypnotically- and environmentally-induced disgust have demonstrated that manipulating participants' level of disgust increases the severity of moral judgement and can lead to a moral judgement in cases exhibiting no

[1] In line with recent work in affective science, I use 'affective' to refer to a range of states, including (but not limited to) emotions, which involve the perception of bodily change, hedonic valence (i.e. felt pleasure or displeasure) and arousal (activation of the autonomic nervous system and endocrine system) (e.g. Russell 2003, Barrett and Bliss-Moreau 2009, cf. Prinz 2004). Following Greene et al. (2004, pp. 397–8), I use 'affect' contrastively with 'cognitive'. I discuss this further in the text.

 © Zarpentine 2016

apparent transgression (Wheatley and Haidt 2005; Schnall et al. 2008). In a phenomenon dubbed *moral dumbfounding*, participants persist in endorsing their emotion-driven moral judgements despite being unable to produce a reasoned justification (Haidt, Bjorkland, and Murphy 2000).

Joshua Greene has developed a similar view (Greene et al. 2001; Greene et al. 2004; Greene 2009). Greene and his colleagues sought to shed light on a puzzling series of widely-discussed cases involving runaway trolleys (Thomson 1976). These studies utilized functional magnetic resonance imaging to reveal differential patterns of neural activation in response to such dilemmas. What these studies reveal is that *personal dilemmas* like the standard *Footbridge Case* (which involve causing direct harm to a particular person that is not the result of a deflection of an existing danger to another individual) triggered areas of the brain associated with emotion and social cognition (e.g. the medial prefrontal cortex and the amygdala). *Impersonal dilemmas*, like the standard *Trolley Case* (which involve no such direct harm), trigger greater activation in regions associated with working memory, abstract reasoning, and other more *cognitive* processes (e.g. the dorsolateral prefrontal cortex) (Greene et al. 2001; Greene et al. 2004).

This work indicates that, while intuitive, affective processes play a key role in moral judgement, in some cases processes of abstract reasoning and cognitive control are also involved. Greene et al. suggest that the important distinction between affective and cognitive processes, respectively, might be drawn in terms of 'representations that have direct motivational force and ... representations that have no direct motivational force on their own, but that can be contingently connected to affective/emotional states that do have such force' (2004, pp. 397–8. See also Kable and Glimcher 2009). Thus, Greene's view combines neurosentimentalism with externalism about processes that are more cognitive.

Adina Roskies (2003) has drawn on research in clinical psychology focused on vmPFC patients to argue against internalism. She argues that individuals with damage to the vmPFC present 'walking counterexamples' to internalism: they demonstrate normal declarative knowledge and reasoning in the moral domain, yet often lack motivation in 'ethically charged situations' (2003, p. 57). The peculiar pattern of dysfunction demonstrated by individuals with damage to the vmPFC has become widely known primarily through the work of Antonio Damasio and his colleagues (e.g. Damasio 1994; Bechara et al. 1997). Such individuals exhibit generally flattened affect and serious decision

 © Zarpentine 2016

making difficulties. Individuals with vmPFC damage exhibit difficulty on the Iowa Gambling Task (IGT), an experimental measure designed to approximate real-world decision making, and, during the task, fail to produce the same physiological responses (skin conductance responses (SCRs)) as normal individuals (Bechara et al. 1994; Bechara et al. 1996; Bechara et al. 1997). Yet, even as they make decisions which predictably lead to personal and financial ruin, they retain their previously acquired moral and social knowledge (Saver and Damasio 1991). Thus, while their capacity to make moral judgements persists following the traumatic brain injury, their decision making difficulties and lack of physiological responses indicate a motivational deficit. Roskies argues that such cases present a counterexample to the view that there is a necessary connection between moral judgement and motivation.

Subsequent studies have found only one systematic effect of vmPFC damage on moral judgement: individuals with vmPFC damage are more likely to endorse actions that will cause personal harm in order to secure greater beneficial consequences (e.g. pushing the large man in the Footbridge Case, in order to save five others) (Koenigs et al. 2007; Ciaramelli et al. 2007; Moretto et al. 2010; Thomas, Croft, and Tranel 2011). It appears that previously acquired social and moral beliefs allow individuals with damage to the vmPFC to make the (statistically) normal judgements in many cases. But responding to personal moral dilemmas depends upon affective processing that is disrupted following vmPFC damage. Without negative affective responses to the thought of pushing a large man in front of a trolley, vmPFC patients make the characteristically utilitarian judgement that it is permissible to push the large man to his death in order to save the lives of five others. This receives confirmation from evidence indicating that normal participants who made more utilitarian judgements in personal moral dilemmas generated fewer SCRs prior to utilitarian judgement, whereas participants who made fewer utilitarian judgements showed higher levels of SCRs prior to utilitarian judgement (Moretto et al. 2010).

Similarly, Jesse Prinz (2007) and Shaun Nichols (2004) have also developed versions of sentimentalism informed by research in developmental psychology and psychopathology. Psychopathy involves a pattern of dysfunction that, although importantly distinct, is instructively similar to that exhibited by patients with damage to the vmPFC. Psychopaths exhibit anti-social behavioural tendencies and affective deficits (particularly with respect to guilt, remorse, and empathy) that

 © Zarpentine 2016

derive from a dysfunction of affective learning mechanisms subserved by the amygdala and vmPFC (Blair, Mitchell, and Blair 2005; Blair 2007). Despite this, many cognitive capacities appear intact. Moreover, while psychopaths demonstrate a (statistically) normal pattern of moral judgement (Cima, Tonnaer, and Hauser 2010; Glenn, Raine, and Schug 2009), it is thought that their understanding of morality is abnormal.[2]

Nichols draws on this research in defending his *sentimental rules* account of moral judgement. On his view, moral judgement requires both the representation of moral norms and an affective mechanism that confers special status on those norms (2004, p. 3). Nichols argues that while psychopaths do acquire representations of moral rules, since their affective deficit disrupts normal understanding of morality and since this cannot be traced to any rational failure, this research provides support for sentimentalism over competing rationalist accounts. Nichols also argues that this research presents a difficulty for internalism. Psychopaths demonstrate an awareness of moral norms without motivation to act accordingly (Nichols 2004, pp. 111–2). And even when a psychopath is described as such, people still tend to agree that the psychopath does understand moral norms (2004, pp. 74–5). Thus, Nichols endorses both claims criticized by G&K.

In contrast, Prinz argues that if there is no emotion-independent way of understanding morality and identifying moral norms, then there is good reason to believe that moral properties are essentially related to emotions (Prinz 2007, p. 47). The research on psychopathy strongly suggests that there is no emotion-independent way to characterize moral norms; thus, he endorses a form of neurosentimentalism according to which emotion is both necessary and sufficient for moral judgement. Since emotion has some motivational component, moral judgement will bring with it some motivational force. Thus, while Prinz does endorse internalism, his sentimentalism is subject to G&K's criticism.

[2] Nichols (2004, Ch. 1) reviews research indicating that psychopaths fail to distinguish moral from conventional violations. Recent studies have challenged the diagnostic value of this task (e.g. Kelly et al. 2007) and questioned whether, when more appropriate stimulus materials are used, psychopaths really do fail to distinguish moral from conventional violations (Aharoni, Sinnott-Armstrong, and Kiehl 2012). Nevertheless, psychopaths do exhibit affective dysfunction and studies demonstrate differential processing of emotional stimuli relative to control subjects (e.g. Kiehl 2006). This suggests the view that psychopaths compensate for affective deficits by engaging more cognitive processing.

 © Zarpentine 2016

In this section, I have discussed the work of several philosophers and psychologists who have argued that affective processes play a key role in moral judgement.[3] Research from psychopathology seems to indicate that, even in cases of affective dysfunction, cognitive processes may sustain moral judgement. Drawing on these results, Nichols and Roskies argue in favour of rationalist externalism. Despite endorsing internalism, Prinz argues for a version of neurosentimentalism. And G&K explicitly address his view arguing that by endorsing the view that emotion is both necessary and sufficient for moral judgement, Prinz's neurosentimentalism is committed to the claim that moral judgement does not require an agent (2010, p. 594). In addition to discussing several views that are subject to G&K's criticism, this section briefly introduced empirical research that helps set the stage for the critical discussion below (§3). In the next section I discuss the criticisms presented by G&K.

2. The agential constraint

Neurosentimentalist views focus on the role of implicit, affective processes in generating moral judgements. Rationalist externalist views focus on the role of reasoning that utilizes explicitly represented rules in moral judgements. Thus, G&K claim that both focus exclusively on synchronic processes occurring at the time of judgement. However, moral decision making, like decision making in general, is 'the exercise of a capacity for agency' (2010, p. 585). By focusing only on processes occurring at the time of action these synchronic accounts of moral judgement ignore the *reflective capacities* that are crucial for moral agency. As a result such theories 'leave the moral agent out of the picture' (2010, p. 606). Thus, there is an *agential constraint* on theories of moral judgement and both neurosentimentalist and rationalist externalist theories fail to meet this constraint.

According to G&K, moral agency requires that an agent be able to view herself as a 'temporally extended entity' (2010, p. 588). This ability, they claim, depends on the psychological capacity for *mental time travel*. Mental time travel is sometimes called episodic future thought (Peters and Büchel 2010), prospection (Gilbert and Wilson 2007), or

[3] Despite having support from a body of research in a range of fields utilizing a variety of experimental methodologies, the claim that emotions play a causal role in moral judgement has been contested (e.g. Huebner, Dwyer, and Hauser 2009; Cima, Tonnaer, and Hauser 2010). Since G&K's criticism does not depend on this kind of scepticism, I will not discuss it further.

 © Zarpentine 2016

simulation (Schacter, Addis, and Buckner 2008). It is the capacity to imagine future events and is particularly helpful in predicting the consequences of actions. Interestingly, the capacity for mental time travel depends upon the hippocampus and medial temporal lobe (MTL) and these same regions are implicated in declarative memory (which includes semantic and episodic memory, i.e. the ability to recall facts and events) (Schacter, Addis, and Buckner 2009). While leaving certain forms of implicit and affective learning intact, damage to this memory system results in *anterograde amnesia*: the inability to form new declarative memories (Corkin 2002; Schmolck et al. 2002; Rosenbaum et al. 2005; Tulving et al. 1988). Such individuals, often with a memory of less than 30 seconds, have been said to live in a 'permanent present' (Barbizet 1970, p. 33, as quoted in Tulving et al. 1988).

According to G&K, mental time travel is both crucial for moral agency and lacking in vmPFC patients. They write: 'It seems likely that the normal case of decision-making involves mental time travel' (2010, p. 600). They continue:

> the capacity for mental time travel is arguably required for *any* demand to be normative for a person—for them to have reasons whose force is independent of immediate stimulus bound responses … the process of becoming an agent capable of engaging in this normative domain [morality] just is the process of learning to transcend the present moment, both cognitively and behaviourally. (2010, p. 602, emphasis in the original)

Thus, these *reflective capacities*, to conceive of oneself as a temporally extended entity and to take certain considerations as normative, are central to moral agency. They are not sufficient for moral agency but they are a 'prerequisite for moral judgement' (2010, p. 606).

There are several ways in which to interpret this agential constraint. A *strong* interpretation might hold that *every* moral judgement involves the exercise of these reflective capacities. Such a constraint is implausibly strong, placing an undue burden on moral agents. Many ordinary cases of moral judgement do not involve the exercise of such reflective capacities. I do not need to rehearse the relevant reasons every time I judge that torture is wrong. Moreover, there is some indication that G&K do not intend to endorse the strong agential constraint.[4]

[4] They write: 'There are many cases of moral judgement which are not well captured by the synchronic focus of most experimental work on moral cognition' (2010, p. 607). This suggests that *some* cases of moral judgement *do* derive from the kinds of processes described by neurosentimentalist and rationalist externalist accounts.

 © Zarpentine 2016

A weaker interpretation holds that moral judgement does not require the exercise of these reflective capacities in every case, only that the individual making the moral judgment has the *ability* to exercise these reflective capacities. Since this is both much more plausible and better aligned with what G&K seem to have in mind, I will focus on this weaker interpretation. The weaker interpretation could even be strengthened somewhat by adding the stipulation that, on at least some occasions, moral agents must exercise their reflective capacities.

According to G&K, the vmPFC plays an important role in 'helping to create a moral agent constituted by the connection between deliberation, emotion, and diachronic selfhood' (2010, p. 590). Indeed, they claim that 'a patient with ventromedial damage cannot perform mental time travel' (2010, p. 600). As I discuss below (§3), the evidence they provide in support of this claim is slim. They rely primarily on the fact that a deficit in mental time travel would explain vmPFC patients' poor performance on the IGT, a condition Bechara et al. (1994) label '"myopia" for the future'.

On the basis of this claim, they draw an important comparison between individuals with vmPFC damage and those suffering from anterograde amnesia. As indicated, anterograde amnesia is caused by damage to the MTL and results in an inability both to form new declarative memories and to engage in mental time travel. However, since accounts of moral judgement offered by rationalist externalists and neurosentimentalists focus only on synchronic processes, they would seem to allow that both amnesiacs and vmPFC patients can be moral agents, despite their apparent inability to exercise the reflective capacities highlighted by G&K as crucial for moral agency. As a result, neurosentimentalists and rationalist externalists face a dilemma: either they must deny the connection between moral judgement and moral agency or accept the implausible claim that amnesiacs and those with vmPFC damage can be moral agents (2010, p. 588). Neither horn, G&K claim, is attractive.

3. Affect, judgement, and the vmPFC

Undoubtedly there is a connection between moral judgement and agency. However, we should leave room for the possibility of practical irrationality; some dissociation between what we judge it best to do and what we actually do is a natural feature of agents like us. Thus, in

 © Zarpentine 2016

evaluating G&K's critique, we must take care not to require too strong a connection. Additionally, it is important to draw attention to an important feature of the affective mechanisms that neurosentimentalists claim play an essential role in moral judgement. G&K describe affective responses as 'phylogenetically ancient motivational impulses' (2010, p. 590). In fact, affective mechanisms are essentially *learning* mechanisms, implementing a very powerful form of gradual, error-based learning (Kable and Glimcher 2009; Glimcher et al. 2009, Pt. IV).[5] Thus, affective responses deriving from these mechanisms are not comparable to reflexes or innate responses; they are produced by mechanisms that have been shaped over the course of ontogeny. This feature of affective mechanisms is neglected by G&K but, as I discuss below, it has important implications for assessing neurosentimentalism.

In this section, I argue that a proper interpretation of the empirical literature on vmPFC damage undermines the comparison that G&K draw between anterograde amnesia and vmPFC damage (§3.1). This blunts the force of the second horn of the dilemma. Moreover, I argue that affective mechanisms are at least partly constitutive of an agent who persists through time (§3.2). This gives a suitably developed neurosentimentalism the resources to provide a plausible account of the connection between judgement and agency. This argument also blunts the force of the first horn of the dilemma. In fact, a better understanding of the role of the vmPFC indicates precisely how affect contributes to the reflective capacities that G&K claim are crucial for agency. Far from posing a problem, focusing on these reflective capacities may actually vindicate a sophisticated neurosentimentalism.

It is quite plausible that anterograde amnesiacs are not moral agents.[6] However, I think it is a mistake to assimilate vmPFC patients to anterograde amnesiacs. To clarify this issue it is helpful to note two conceptual points. First, we must remember that there are different *varieties* of agency. It is plausible that children and many nonhuman animals are *agents*, in a basic sense; they certainly have the capacity to

[5] This interpretation of the empirical literature on these mechanisms is not without its critics. For example, Berridge (2012) argues that processing in the mid-brain dopamine system represents not *learning* but what he calls 'incentive salience'. For different reasons, Huebner et al. (2009) object that emotion may simply be an effect of the computations which issue in judgement rather than constitutive of those computations. I cannot respond to these criticisms here.

[6] They certainly do not seem to meet the conditions required for moral responsibility. For an interesting discussion of this issue in the context of the film *Memento*, see McKenna (2009).

 © Zarpentine 2016

act (cf. Mele 2003, Ch. 10). We can then distinguish more robust varieties of agency, for example *moral agency*, *rational agency*, or a kind of *social agency*, each requiring a set of capacities beyond basic agency. We can understand G&K as claiming that moral agency requires certain reflective capacities, including the ability to see oneself as a temporally extended entity and to take certain considerations as normative.

Second, it is important to note that different varieties of agency need not be all-or-nothing capacities; they may be impaired to a greater or lesser extent. As children develop, they gradually acquire more robust forms of agency. Acute brain damage or other pathologies can produce impairments in certain capacities that need not completely undermine a particular kind of agency. For example, social agency might require certain social skills like the ability to easily read others' facial expressions and respond appropriately. Individuals with an Autism Spectrum Disorder might exhibit varying degrees of impairment with respect to such capacities. And in certain circumstances, social agency may be required for moral agency. In other circumstances deficits in social agency may not impair moral agency. We might think that individuals who exhibit difficulty with social agency are partially impaired, but not completely incapacitated, with respect to moral agency.

In light of these two conceptual points, it is important to note the possibility that vmPFC patients may exhibit deficits with respect to certain capacities that play a role in moral agency. These may partially disrupt moral agency without entirely compromising it. In fact, I think this is the best way to think about the effect of vmPFC damage on moral agency. If, contrary to G&K, vmPFC patients do not lack moral agency altogether, then the neurosentimentalist pays no theoretical cost in claiming that such individuals can be moral agents. Providing support for the antecedent requires critical attention to the literature G&K cite in supporting their claim.

3.1 The function of the vmPFC

As I discussed above (§2), G&K claim that moral agency requires the capacity for mental time travel but that vmPFC patients lack this capacity. In contrast, I argue that the vmPFC is not crucial for mental time travel per se, but for affective response to such simulation and the integration of this information in decision making.

It is helpful to begin by drawing some distinctions. As noted (§2), damage to the MTL results in an inability to form new declarative

 © Zarpentine 2016

memories, but leaves intact various affective learning mechanisms. Let us distinguish between judgements based on remembered preferences and judgements that rely (in part) on mental time travel. Judgements of the latter kind depend upon what can be called *dynamic evaluation*. Dynamic evaluation contrasts with memory-based preference judgements in being 'an ongoing, dynamic assessment of relative value' (Fellows and Farah 2007, p. 2673). Because remembering the past and imagining the future both depend upon the MTL, both remembered preferences and dynamic evaluation in response to mental time travel will depend upon this area. However, dynamic evaluation can also be engaged *in situ*, in response to perception of the immediate environment. Most importantly, in addition to simulation or perception, dynamic evaluation draws upon affective capacities.

Consider a more concrete example. When ordering at a restaurant I frequent, I may glance at the menu, recognize the seared ahi tuna as an entrée I have enjoyed recently and order that again. Here, I simply rely on declarative memory to make a decision. In contrast, I may be in the mood to try something different. In this case, I may engage in dynamic evaluation: imagining how the flavours of different options will combine and noting my affective response. Do I think I will enjoy a blue cheese-ale sauce with steak? Would mushroom-liver risotto be better? In this second case, I am engaged in dynamic evaluation in response to mental time travel. I am using my affective responses to mental time travel to guide my decision. On the other hand, if I am deciding what to eat at a buffet, I can depend on the actual sight and smell of the food without needing to use my imagination much.

I claim that damage to the vmPFC leads to a dysfunction in dynamic evaluation. Dynamic evaluation can involve response to the immediate environment or response to mental time travel. Contrary to G&K's interpretation, vmPFC patients are not impaired with respect to mental time travel per se. Rather, their problem with dynamic evaluation results from dysfunction of the mechanisms that generate affective responses. Indeed, they appear to be equally impaired in their ability to engage in dynamic evaluation *in situ* or in response to mental time travel. And this explains their problems with emotional regulation and decision making. For example, while they exhibit generally flattened affect, individuals with vmPFC damage are prone to bursts of uncontrolled emotion: one patient is described as berating her almost 90-year-old mother in a clinic waiting room, when her mother could not produce an appointment slip quickly enough (Koenigs and Tranel 2007, p. 954). In this case, dynamic evaluation

 © Zarpentine 2016

would allow an individual to 'dial down' her emotional response; its absence leads to an uncharacteristic explosion of anger. Failure of dynamic evaluation *in situ* also explains the poor performance of vmPFC patients on the IGT and their blatantly poor decision making.

The failure of dynamic evaluation in response to mental time travel can provide an explanation of vmPFC patients' idiosyncratic pattern of moral judgement. They exhibit atypical moral judgements on precisely those cases that require dynamic evaluation in response to mental time travel. Just as I can depend on my remembered preferences when ordering at my favourite restaurant, in many cases vmPFC patients can rely on moral rules that they have learned and stored in declarative memory. In personal moral dilemmas, however, no previously learned moral rule straightforwardly applies. When individuals must weigh directly causing harm against beneficial consequences, dynamic evaluation plays a crucial role. When we consider the Footbridge Case, we may imagine pushing a large man in front of a trolley in order to save five lives and use our affective response to this imagining to guide our decision making. For normal individuals, this thought triggers a negative affective response leading most to judge the action to be impermissible. Since vmPFC patients lack these responses, their judgements are driven more by abstract reasoning focused on the beneficial consequences and so they exhibit higher rates of characteristically utilitarian judgement.

Two additional sources of evidence that support my interpretation are worth noting. First, there is little evidence that vmPFC patients have a deficit in mental time travel per se. Damasio's patient, E.V.R., can imagine future possibilities; he showed normal performance on tasks designed to measure the ability to think of alternative options in a given situation, reflect on the future consequences of actions, engage in means-end reasoning, and foresee the outcome of social situations (Saver and Damasio 1991). Second, there is no evidence that vmPFC patients have memory problems, problems which would be expected based on the dependence of both declarative memory and mental time travel on the MTL. E.V.R. can explain that he had a meningioma and describe the events that followed its removal. As Damasio reports, he 'was able to recount the tragedy of his life with a detachment that was out of step with the magnitude of the events…' (Damasio 1994, p. 44). What is lacking is any affective response to these memories. Damasio continues: 'He was not inhibiting the expression of internal emotional resonance or hushing inner turmoil. He simply did not have any turmoil to hush' (1994, p. 44). Thus, vmPFC patients lack any affective

 © Zarpentine 2016

response to imagined future events or remembered past events. This is in stark contrast to individuals with anterograde amnesia, who are equally unable to imagine future events or to remember past events subsequent to the damage.

Indeed, it is the preserved recall of previously learned social and moral knowledge that makes vmPFC patients so puzzling. Such individuals exhibit declarative memory of the very rules they violate (e.g. Beer et al. 2006). Toward the end of a session of the IGT, some vmPFC patients are actually able to report on which decks are good and which are bad—even as they continue to draw from the bad decks (Bechara et al. 1996; Bechara et al. 1997)! All of this points to the conclusion that the failure of dynamic evaluation in vmPFC patients rests not with an impairment of the capacity for mental time travel, but with a dysfunction of the mechanisms that generate affective responses to such imaginings.[7]

In support of their own interpretation, G&K provide only limited evidence. They cite one case study (Levine et al. 1999) of an individual with 'amnesia and ventromedial damage with deficient mental time travel … *and* decision making deficits … which indicate a loss of agency' (2010, p. 600, emphasis in the original). The subject of this study, M.L., is reported to have *retrograde* amnesia (i.e. impaired memory of events previous to the injury) and partially preserved anterograde memory (Levine et al. 1998). A more recent discussion of this case suggests a dissociation between semantic and episodic memory and that M.L.'s deficits in mental time travel might, in fact, be due to M.L.'s retrograde amnesia (Levine et al. 2009). This case does not provide very compelling evidence that the vmPFC is necessary for mental time travel.

G&K also cite a study indicating poor performance on the IGT in amnesia patients (Gutbrod et al. 2006). It is likely that several processes are involved in the IGT (e.g. Lawrence et al. 2009). That declarative memory (subserved by the MTL) is required for performance on the IGT should not be completely surprising. It does not, however, provide evidence that vmPFC patients also have a deficit in mental time travel. Rather, it simply indicates what we already know: performance on the IGT does not depend *only* on the vmPFC.

[7] This is consistent with a wide range of research on the function of the vmPFC and neighbouring regions of the medial PFC and orbitofrontal cortex (e.g. Fellows and Farah 2003, 2005a, 2007; Kable and Glimcher 2009; Kringelbach and Berridge 2009; Reimann and Bechara 2010; Rolls 2005, Ch.4).

 © Zarpentine 2016

Another study, which G&K do not cite, could be taken to provide evidence for their interpretation. Fellows and Farah (2005b) found foreshortened future time perspective following vmPFC damage, indicating that when vmPFC patients imagine the future, they do not look as far ahead as normal individuals. This could be taken to indicate a deficit in mental time travel. But it could also be due simply to the affective dysfunction that disrupts dynamic evaluation. Affective responses heighten the salience of thoughts and perceptions, increasing the likelihood of reporting. Lacking such responses, vmPFC patients may simply be less likely to report their distant future imaginings.

Recently, Pascal Boyer put forward the view that mental time travel serves as a 'brake' on impulsive or '"myopic" decision making' (2008, p. 219). If correct, this view might provide support for G&K's hypothesis that vmPFC patients are impaired with respect to mental time travel. However, a more recent study (which includes Boyer as a co-author) found evidence that anterograde amnesiacs exhibit typical discounting of future rewards, casting doubt on this view (Kwan et al. 2012). And this research is consistent with the view argued for here that damage to the vmPFC undermines dynamic evaluation.[8]

Thus, there are good reasons to reject G&K's alternative interpretation of the research on vmPFC damage. On the view I have supported, vmPFC patients are not impaired with respect to mental time travel and their affective capacities are not completely compromised. As a result, they seem to meet the conditions that G&K claim are necessary for moral agency: to be able to see oneself as a temporally extended entity and to take certain considerations as normative. Of course, it is plausible that a deficit in dynamic evaluation does impair moral agency to some extent, especially in certain kinds of situations; but it does not render them completely incapacitated with respect to moral agency.[9]

[8] Kwan et al. examined temporal discounting using a standard task requiring participants to choose between an immediate reward and a larger future reward. It should be noted that there is a difference between making preference judgements about present and future rewards and successfully resisting the temptation to forgo greater future rewards in favour of smaller immediate ones. It is possible that dynamic evaluation in response to mental time travel is much more important for the latter. And this is consistent with vmPFC patients' characteristic myopic decision-making. Thanks to an anonymous referee for bringing my attention to the research discussed in this paragraph.

[9] It might be argued that the weak interpretation of the agential constraint, which I ascribed to G&K above, is too weak and that it can be strengthened enough to rule out pathological cases. Apart from appearing ad hoc, this strategy faces a serious problem: there is extensive diversity within moral psychology. Several studies indicate significant overlap between normal and pathological groups. To take two examples: (1) some otherwise normal

 © Zarpentine 2016

3.2 Affect and the constitution of agency

I now want to discuss further the feature of affective mechanisms I drew attention to above: in contrast to G&K's characterization of affective responses as 'phylogenetically ancient motivational impulses' (2010, p. 590), such responses actually derive from sophisticated learning mechanisms that are modified over the course of an individual's development and are particularly sensitive to the kind of training that occurs in moral education (e.g. Hoffman 1994). Gilbert Ryle (1958), in his discussion of the absurdity of claiming to have forgotten the difference between right and wrong, highlights the role inculcation during upbringing plays in our coming to appreciate this difference. When one's moral sensibilities weaken or change it is not a case of forgetting something but, Ryle claims, a case of there being a change in the person. Our knowledge of right and wrong is central to our identity as agents. And affective mechanisms are precisely the kind of learning mechanisms that are shaped by training and inculcation during development. Thus, I think there is good reason to think of affective mechanisms as (at least partially) *constitutive of one's agency* and to think of affective responses as *expressions* of that agency and not as some 'ancient motivational impulses' that impose themselves as an alien force on an agent.

One way of understanding what makes an affective response an expression of an individual's agency, rather than something external that simply happens to the agent, involves the notion of internality[10] or identification (e.g. Frankfurt 1976). To borrow from an example of Frankfurt's, at hearing of a good friend's recent accomplishment, I may feel a desire to pay her a compliment. At the same time, I may also experience a sudden jealous desire to harm her by belittling the achievement. This latter desire may be experienced as an unwelcome imposition upon me and I may do my best to eliminate it. In this case,

---

individuals exhibited the same patterns of moral judgements and physiological responses as vmPFC patients (Moretto et al. 2010) and (2) a subset of addicts as well as a subset of individuals in normal control groups exhibit behavioural and physiological performance on the IGT similar to that of vmPFC patients (Bechara 2005). While some of these examples involve individuals with substance abuse problems, others appear to suffer no abnormalities besides a penchant for utilitarianism—which can hardly be considered a pathology. In light of this diversity, it is implausible that the agential constraint can be strengthened in a way that excludes pathological individuals without also excluding some otherwise normal individuals.

[10] This notion of internality, which describes the relation between an agent and her attitudes, should not be confused with the notion of 'internalism' discussed in §1, which describes the relation between motivation and moral judgement. Thanks to an anonymous reviewer and the editor of this journal for encouraging me to further develop the argument in this section.

 © Zarpentine 2016

the first desire is internal while the second is external. Distinguishing internal from external attitudes is, of course, difficult and there is no guarantee that one's own feelings of identification will be completely reliable. Still, it might be thought that, in contrast to basic agency, the capacity for *internal* attitudes is essential for more sophisticated forms of agency. Indeed, it is sometimes thought that the possession of internal attitudes is a unique characteristic of human agents.

Many standard accounts of internality require a certain kind of reflectiveness.[11] In particular, an emotion or attitude is internal only if an agent possesses a reflexive attitude of endorsement toward that emotion or attitude. On such views, internality requires the capacity to make one's motivational attitudes the target of reflection and endorsement. Though they do not directly discuss this literature, G&K's agential constraint seems to be in the spirit of these views. As I discussed above, they claim that moral judgement is the exercise of a capacity for agency and that agency requires certain reflective capacities, including the ability to conceive of oneself as a temporally extended being and to take certain considerations as normative—plausibly capacities that allow an agent to reflectively endorse their motivational attitudes. Thus, if we assume that moral agency requires the capacity for internal attitudes and that internal attitudes require certain reflective capacities, then we would have some reason to endorse the agential constraint as G&K present it.

Recently, Agnieszka Jaworska (2007) has criticized such accounts of internality. She argues that they fail to account for the internal attitudes of 'marginal agents' like children or Alzheimer's patients. Though their moral agency may be underdeveloped or impaired in some way, two-year-old children and Alzheimer's patients both seem capable of *caring*, which Jaworska considers an 'inherently internal' attitude (2007, p. 532). Thus, despite the fact that they may lack (at least to some extent) the relevant reflective capacities or fail to possess the appropriate reflexive attitudes, it is very plausible that such agents are nevertheless capable of possessing internal attitudes. This poses a problem for standard accounts of internality.

Jaworska goes on to develop accounts of caring and internality that can handle these cases of marginal agents. She concludes:

> any attitude, reflexive or not, has the right kind of authority to speak for the agent, so long as it is part of its function to support the psychological

[11] For examples, see Frankfurt (1971) and Bratman (2000). For further discussion, see Jaworska (2007, pp. 541–9).

 © Zarpentine 2016

> continuities and connections that constitute the agent's identify and cohesion over time. (2007, p. 552)

Contrasting 'episodes of emotional experience' with 'emotions', which she describes as 'a more enduring state comprised of various interrelated emotional episodes, and dispositions to experience subsequent emotional episodes, unfolding intermittently over time' (2007, p. 552), Jaworska claims that emotions can have the structure necessary to realize the complex continuities and connections required to sustain internal attitudes.[12] Thus, even in the absence of reflective capacities, marginal agents can satisfy the conditions for the possession of internal attitudes. Rejecting G&K's view of emotion as akin to reflexes and endorsing the account of affective mechanisms as sophisticated learning mechanisms shaped over the course of development can both explain how emotions come to have the requisite structure and how these affective mechanisms can play a role in the constitution of an agent that persists through time.[13] In failing to appreciate this diachronic aspect of affect, G&K neglect an important way in which neurosentimentalists can explain both the role affect plays in agency and the connection between judgement and agency.

Further support comes from attending to the processes of development and degeneration of moral agency. In children, moral agency emerges over time—it does not suddenly appear with the capacity for reflexive attitudes. This makes perfect sense on the view that the development of affective capacities plays a significant role in the formation of moral agency. The powerful learning mechanisms that constitute these affective mechanisms develop gradually and are sensitive to the training and inculcation that occur throughout childhood. As a result, moral agency begins to emerge even before the development of more cognitive reflective capacities. In the case of Alzheimer's patients, we see this process in reverse. Higher cognitive

[12] Incidentally, Prinz draws a similar distinction between emotions and sentiments, the latter of which he describes as 'a disposition whose occurrent manifestations … are emotions' (2007, p. 84). Care, according to Prinz, is a sentiment: it is 'a disposition to experience different emotions' (2007, p. 84). In different circumstances, caring for something may give rise to different emotions: you may feel happy that you have it, or afraid that you will lose it, envious that someone else has it or hopeful that you may get it, and so on. Together with Jaworska's account of internality, such neurosentimentalist accounts can meet G&K's challenge by explaining how emotions or sentiments are (at least partially) constitutive of agency. Indeed, Prinz claims moral sentiments 'are the backbone of morality' (2007, p. 86).

[13] This is not to deny that some emotional responses are external. Nevertheless, when we articulate a moral judgement that derives from an emotional response with the appropriate continuities and connections, that judgement is an expression of our identity as agents.

 © Zarpentine 2016

skills degrade, leaving affective capacities largely intact and allowing these individuals to continue to exhibit internal attitudes (e.g. caring) despite the loss of important reflective capacities.

The dramatic shifts in perceived identity of individuals who have suffered acute damage to the vmPFC further illustrates how these mechanisms are partially constitutive of agency. Consider Phineas Gage, perhaps the most well-known individual to have suffered damage to the vmPFC. Following his injury, his physician reports: 'his mind was radically changed, so decidedly that his friends and acquaintances said he was "no longer Gage"' (Harlow 1868, pp. 339–40). This oft-quoted passage reflects exactly what we would expect on the view that such affective capacities are partially constitutive of agency: damage to these mechanisms leads to striking changes in an individual's agential identity, despite the preservation of higher cognitive skills.

Jaworska's account helps explains how, in the absence of the higher cognitive capacities necessary for reflection, an individual may nevertheless possess internal attitudes through the possession of affective mechanisms that exhibit the requisite structure. However, it bears mentioning that affect can also play a role in more reflective processes. As G&K put it: 'Moral reflection, planning, and choice … engages our sense of self and our capacity to see ourselves, and others, and the world in which we find ourselves, diachronically' (2010, p. 607). It is very plausible that when we engage in this sort of reflection, we often rely on what I've called dynamic evaluation in response to simulation. This does depend on mental time travel; thus, G&K are right to highlight the importance of this capacity for moral agency. However, as I've argued, this process also depends upon affective mechanisms, especially the vmPFC. If affective mechanisms play a role in both moral judgement and the exercise of reflective capacities, then correcting G&K's misinterpretation of the research on the function of the vmPFC provides the neurosentimentalist with the resources to better explain the connection between judgement and agency. Properly interpreting this research turns out to be fundamental to evaluating G&K's criticism of neurosentimentalism.

G&K conclude by reiterating the diachronic aspects of moral agency and by calling for greater attention to the 'processes which create and sustain agency' (2010, p. 607). If the argument put forward here is correct, such a project cannot proceed without appreciating the crucial role that affective mechanisms play in these processes. And this, it might be thought, actually provides support for neurosentimentalism.

 © Zarpentine 2016

4. Conclusion

Gerrans and Kennett claim that moral judgement is the exercise of a capacity for agency and that moral agency requires the capacity to see oneself as a temporally extended entity and to take certain considerations as normative. Crucial to these later capacities is the ability to engage in mental time travel. They claim that damage to the vmPFC disrupts the ability to engage in mental time travel and that therefore vmPFC patients cannot be moral agents. They argue that this poses a problem for those who draw on neuroscientific and psychological research to support a form of moral sentimentalism: because they focus only on synchronic processes occurring at the time of moral judgement, a consequence of such views is that vmPFC patients can be moral agents.

I have argued that G&K ignore an important diachronic aspect of affective mechanisms and misinterpret the empirical research on the function of the vmPFC. G&K claim that the vmPFC plays a key role in moral agency. This is undoubtedly correct. However, the ability to see oneself as a temporally extended entity and to take certain considerations as normative does not depend only on the vmPFC, but is explained by the complex interplay of a variety of capacities, both affective and cognitive, including mental time travel and dynamic evaluation. Careful attention to the empirical research on the function of the vmPFC arguably provides support for neurosentimentalism.

According to the interpretation I have defended, vmPFC patients meet G&K's conditions for moral agency after all, although it is quite plausible that their agency is somewhat impaired. I have denied neither the connection between moral judgement and agency nor the importance of reflective capacities for the latter. However, given the diachronic aspects of affective mechanisms, neurosentimentalism has the resources to offer a robust account of both how affective mechanisms are partially constitutive of an agent that persists through time and how moral judgements that issue from these mechanisms can be expressions of such agency. Sophisticated forms of neurosentimentalism can easily accommodate the agential constraint without sacrificing the centrality of affective processes to moral thought and agency.[14]

[14] Thanks to three anonymous referees and the editorial board of this journal for very helpful feedback. Thanks also to Alfred Mele and David McNaughton for feedback on earlier versions of this paper.

 © Zarpentine 2016

References

Aharoni, Eyal, Walter Sinnott-Armstrong, and Kent A. Kiehl 2012: 'Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/conventional Distinction'. *Journal of Abnormal Psychology*, 121, pp. 484–97.

Barbizet, Jacques 1970: *Human Memory and Its Pathology*. San Francisco, CA: W. H. Freeman.

Barrett, Lisa F., and Eliza Bliss-Moreau 2009: 'Affect as Psychological Primitive'. In Zanna 2009, pp. 167–218.

Bechara, Antoine 2005: 'Decision Making, Impulse Control and Loss of Willpower to Resist Drugs: A Neurocognitive Perspective'. *Nature Neuroscience*, 8, pp. 1458–63.

Bechara, Antoine, Antonio R. Damasio, Hanna Damasio, and Steven W. Anderson 1994: 'Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex'. *Cognition*, 50, pp. 7–15.

Bechara, Antoine, Hanna Damasio, Daniel Tranel, and Antonio R. Damasio 1997: 'Deciding Advantageously Before Knowing the Advantageous Strategy'. *Science*, 275, pp. 1293–5.

Bechara, Antoine, Daniel Tranel, Hanna Damasio, and Antonio R. Damasio 1996: 'Failure to Respond Autonomically to Anticipated Future Outcomes Following Damage to Prefrontal Cortex'. *Cerebral Cortex*, 6, pp. 215–25.

Beer, Jennifer S., Oliver P. John, Donatella Scabini, and Robert T. Knight 2006: 'Orbitofrontal Cortex and Social Behavior: Integrating Self-Monitoring and Emotion-Cognition Interactions'. *Journal of Cognitive Neuroscience*, 18, pp. 871–9.

Berridge, Kent C. 2012: 'From Prediction Error to Incentive Salience: Mesolimbic Computation of Reward Motivation'. *European Journal of Neuroscience*, 35, pp. 1124–43.

Blair, R. James 2007: 'The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy'. *Trends in Cognitive Sciences*, 11, pp. 387–92.

Blair, R. James, D.G.V. Mitchell, and K.S. Blair 2005: *The Psychopath: Emotion and the Brain*. Malden, MA: Wiley-Blackwell.

Boyer, Pascal 2008: 'Evolutionary Economics of Mental Time Travel?' *Trends in Cognitive Sciences*, 12, pp. 219–24.

Bratman, Michael E. 2000: 'Reflection, Planning, and Temporally Extended Agency'. *The Philosophical Review*, 109, pp. 35–61.

 © Zarpentine 2016

Ciaramelli, E., M. Muccioli, E. Làdavas, and G. di Pellegrino 2007: 'Selective Deficit in Personal Moral Judgment Following Damage to Ventromedial Prefrontal Cortex'. *Social Cognitive and Affective Neuroscience*, 2, pp. 84–92.

Cima, M., F. Tonnaer, and Marc D. Hauser 2010: 'Psychopaths Know Right from Wrong but Don't Care'. *Social Cognitive and Affective Neuroscience*, 5, pp. 59–67.

Corkin, Suzanne. 2002. 'What's New with the Amnesic Patient H.M.?' *Nature Reviews Neuroscience*, 3, pp. 153–60.

Damasio, Antonio R. 1994: *Descartes' Error*. New York: Putnam.

Fellows, Lesley K., and Martha J. Farah 2003: 'Ventromedial Frontal Cortex Mediates Affective Shifting in Humans: Evidence from a Reversal Learning Paradigm'. *Brain*, 126, pp. 1830–7.

—— 2005a: 'Different Underlying Impairments in Decision-Making Following Ventromedial and Dorsolateral Frontal Lobe Damage in Humans'. *Cerebral Cortex*, 15, pp. 58–63.

—— 2005b: 'Dissociable Elements of Human Foresight: A Role for the Ventromedial Frontal Lobes in Framing the Future, but Not in Discounting Future Rewards'. *Neuropsychologia*, 43, pp. 1214–21.

—— 2007: 'The Role of Ventromedial Prefrontal Cortex in Decision Making: Judgment under Uncertainty or Judgment Per Se?' *Cerebral Cortex*, 17, pp. 2669–2674.

Frankfurt, Harry 1976: 'Identification and Externality'. In Rorty 1976, pp. 239–51.

Fellows, Lesley K., and Martha J. Farah 1971: 'Freedom of the Will and the Concept of a Person'. *The Journal of Philosophy*, 68, pp. 5–20.

Gazzaniga, Michael 2009: *The Cognitive Neurosciences*, 4th edition. Cambridge, MA: MIT Press.

Gerrans, Philip, and Jeanette Kennett 2010: 'Neurosentimentalism and Moral Agency'. *Mind*, 119, pp. 585–614.

Gilbert, Daniel T., and Timothy D. Wilson 2007: 'Prospection: Experiencing the Future'. *Science*, 317, pp. 1351–4.

Glenn, A.L., A. Raine, and R.A. Schug 2009: 'The Neural Correlates of Moral Decision-Making in Psychopathy'. *Molecular Psychiatry*, 14, pp. 5–6.

Glimcher, Paul W., Colin Camerer, Russell A. Poldrack, and Ernst Fehr 2009: *Neuroeconomics: Decision Making and the Brain*. New York: Academic Press.

Greene, Joshua D. 2009: 'The Cognitive Neuroscience of Moral Judgment'. In Gazzaniga 2009, pp. 987–1000.

 © Zarpentine 2016

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen 2004: 'The Neural Bases of Cognitive Conflict and Control in Moral Judgment'. *Neuron*, 44, pp. 389–400.

Greene, J.D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen 2001: 'An fMRI Investigation of Emotional Engagement in Moral Judgment'. *Science*, 293, pp. 2105–8.

Gutbrod, K., C. Krouzel, H. Hofer, R. Müri, W. Perrig, and R. Ptak 2006: 'Decision-Making in Amnesia: Do Advantageous Decisions Require Conscious Knowledge of Previous Behavioural Choices?' *Neuropsychologia*, 44, pp. 1315–24.

Haidt, Jonathan 2001: 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment'. *Psychological Review*, 108, pp. 814–34.

Haidt, Jonathan, F. Bjorkland, and S. Murphy 2000: *Moral Dumbfounding: When Intuition Finds No Reason*. University of Virginia: Unpublished manuscript.

Harlow, John. M. 1868: 'Recovery from the Passage of an Iron Bar through the Head'. *Publications of the Massachusetts Medical Society*, 2, pp. 327–47.

Hoffman, Martin L. 1994: 'Discipline and Internalization'. *Developmental Psychology*, 30, pp. 26–28.

Huebner, Bryce, Susan Dwyer, and Marc D. Hauser 2009: 'The Role of Emotion in Moral Psychology'. *Trends in Cognitive Sciences*, 13, pp. 1–6.

Jaworska, Agnieszka 2007: 'Caring and Internality'. *Philosophy and Phenomenological Research*, 74, pp. 529–68.

Kable, Joseph W., and Paul W. Glimcher 2009: 'The Neurobiology of Decision: Consensus and Controversy'. *Neuron*, 63, pp. 733–45.

Kania, Andrew (ed.) 2009: *Memento*. New York: Routledge.

Kelly, Daniel, Stephen Stich, Kevin J. Haley, Serena J. Eng, and Daniel M.T. Fessler 2007: 'Harm, Affect, and the Moral/Conventional Distinction'. *Mind & Language*, 22, pp. 117–31.

Kiehl, Kent A. 2006: 'A Cognitive Neuroscience Perspective on Psychopathy: Evidence for Paralimbic System Dysfunction'. *Psychiatry Research*, 142, pp. 107–28.

Koenigs, Michael, and Daniel Tranel 2007: 'Irrational Economic Decision-Making after Ventromedial Prefrontal Damage: Evidence from the Ultimatum Game'. *Journal of Neuroscience*, 27, pp. 951–56.

 © Zarpentine 2016

Koenigs, Michael, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc D. Hauser, and Antonio R. Damasio 2007: 'Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements'. *Nature*, 446, pp. 908–11.

Kringelbach, Morten L., and Kent C. Berridge 2009: 'Towards a Functional Neuroanatomy of Pleasure and Happiness'. *Trends in Cognitive Sciences*, 13, pp. 479–87.

Kwan, Donna, Carl F. Craver, Leonard Green, Joel Myerson, Pascal Boyer, and R. Shayna Rosenbaum 2012: 'Future Decision-Making without Episodic Mental Time Travel'. *Hippocampus*, 22, pp. 1215–19.

Lawrence, N.S., F. Jollant, O. O'Daly, F. Zelaya, and M.L. Phillips 2009: 'Distinct Roles of Prefrontal Cortical Subregions in the Iowa Gambling Task'. *Cerebral Cortex*, 19, pp. 1134–43.

Levine, B., S.E. Black, R. Cabeza, M. Sinden, A.R. Mcintosh, J.P. Toth, E. Tulving, and D.T. Stuss 1998: 'Episodic Memory and the Self in a Case of Isolated Retrograde Amnesia'. *Brain*, 121, pp. 1951–73.

Levine, B., M. Freedman, D. Dawson, S. Black, and D.T. Stuss 1999: 'Ventral Frontal Contribution to Self-Regulation: Convergence of Episodic Memory and Inhibition'. *Neurocase: The Neural Basis of Cognition*, 5, pp. 263–75.

Levine, B., E. Svoboda, G.R. Turner, M. Mandic, and A. Mackey 2009: 'Behavioral and Functional Neuroanatomical Correlates of Anterograde Autobiographical Memory in Isolated Retrograde Amnesic Patient M.*L*'. *Neuropsychologia*, 47, pp. 2188–96.

McKenna, Michael 2009: 'Moral Monster or Responsible Person? Memento's Leonard as a Case Study in Defective Agency'. In Kania 2009, pp. 23–43.

Melden, A.I. (ed.) 1958: *Essays in Moral Philosophy*. Seattle, WA: University of Washington Press.

Mele, Alfred R. 2003: *Motivation and Agency*. New York: Oxford University Press.

Moretto, G., E. Ladavas, F. Mattioli, and G. di Pellegrino 2010: 'A Psychophysiological Investigation of Moral Judgment after Ventromedial Prefrontal Damage'. *Journal of Cognitive Neuroscience*, 22, pp. 1888–99.

Nichols, Shaun 2004: *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.

 © Zarpentine 2016

Peters, J., and C. Büchel 2010: 'Episodic Future Thinking Reduces Reward Delay Discounting through an Enhancement of Prefrontal-Mediotemporal Interactions'. *Neuron*, 66, pp. 138–48.

Prinz, Jesse 2004: *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.

——2007: *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Reimann, Martin, and Antoine Bechara 2010: 'The Somatic Marker Framework as a Neurological Theory of Decision-Making: Review, Conceptual Comparisons, and Future Neuroeconomics Research'. *Journal of Economic Psychology*, 31, pp. 767–76.

Rolls, E.T. 2005: *Emotion Explained*. Oxford: Oxford University Press.

Rorty, Amelie O. (ed.) 1976: *The Identities of Persons*. Berkeley, CA: University of California Press.

Rosenbaum, R.S., S. Köhler, D.L. Schacter, M. Moscovitch, R. Westmacott, S.E. Black, F. Gao, and E. Tulving 2005: 'The Case of K.C.: Contributions of a Memory-Impaired Person to Memory Theory'. *Neuropsychologia*, 43, pp. 989–1021.

Roskies, Adina 2003: 'Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy"'. *Philosophical Psychology*, 16, pp. 51–66.

Russell, J.A. 2003: 'Core Affect and the Psychological Construction of Emotion'. *Psychological Review*, 110, pp. 145–72.

Ryle, Gilbert 1958: 'On Forgetting the Difference between Right and Wrong'. In Melden 1958, pp. 147–59.

Saver, J.L., and A.R. Damasio 1991: 'Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy due to Ventromedial Frontal Damage'. *Neuropsychologia*, 29, pp. 1241–49.

Schacter, D.L., D.R. Addis, and R.L. Buckner 2008: 'Episodic Simulation of Future Events: Concepts, Data, and Applications'. *Annals of the New York Academy of Sciences*, 1124, pp. 39–60.

——2009: 'Constructive Memory and the Simulation of Future Events'. In Gazzaniga 2009, pp. 751–62.

Schmolck, H., E.A. Kensinger, S. Corkin, and L.R. Squire 2002: 'Semantic Knowledge in Patient H.M. and Other Patients with Bilateral Medial and Lateral Temporal Lobe Lesions'. *Hippocampus*, 12, pp. 520–33.

 © Zarpentine 2016

Schnall, S., J. Haidt, G.L. Clore, and A. Jordan 2008: 'Disgust as Embodied Moral Judgment'. *Personality and Social Psychology Bulletin*, 34, pp. 1096–1109.

Thomas, B.C., K.E. Croft, and D. Tranel 2011: 'Harming Kin to Save Strangers: Further Evidence for Abnormally Utilitarian Moral Judgments after Ventromedial Prefrontal Damage'. *Journal of Cognitive Neuroscience*, 23, pp. 1–11.

Thomson, J.J. 1976: 'Killing, Letting Die and the Trolley Problem'. *The Monist* 59, pp. 204–17.

Tulving, E., D.L. Schacter, D.R. McLachlan, and M. Moscovitch 1988: 'Priming of Semantic Autobiographical Knowledge: A Case Study of Retrograde Amnesia'. *Brain and Cognition*, 8, pp. 3–20.

Wheatley, Thalia, and Jonathan. Haidt 2005: 'Hypnotic Disgust Makes Moral Judgments More Severe'. *Psychological Science*, 16, pp. 780–4.

Zanna, M.P. 2009: *Advances in Experimental Social Psychology*, vol. 41. Burlington, VT: Academic Press.

© Zarpentine 2016

Mental Time Travel, Dynamic Evaluation, and Moral Agency

PHILIP GERRANS
University of Adelaide
philip.gerrans@adelaide.edu.au

JEANETTE KENNETT
Macquarie University
jeanette.kennett@mq.edu.au

Mental time travel is the ability to simulate alternative pasts and futures. It is often described as the ability to project a sense of self in the service of diachronic agency. It requires not only semantic representation but affective sampling of alternative futures. If people lose this ability for affective sampling their sense of self is diminished. They have less of a self to project hence are compromised as agents. If they cannot "feel the future" they cannot imaginatively inhabit it and hence their agency is compromised. The extent of such losses and consequent impairments to moral agency can be matters of degree.

1. Episodic and autobiographical mental time travel

In 'Neurosentimentalism and Moral Agency' (2010) we argued that:

(i) moral agency is essential for moral judgement

(ii) mental time travel is essential for moral agency

(iii) ventromedial patients have impaired mental time travel

(iv) they have this in virtue of their affective deficits

It follows that ventromedial patients have impaired moral agency and hence impaired moral judgement. The basic idea is that in order to make moral judgements a person has to be able, not only to apply a rule or be motivated appropriately by tacit affective responses, but to mentally represent her*self* in morally relevant future scenarios. We argued that ventromedial patients lose this crucial ability to project themselves in time because they lose the ability to integrate affective responses with representations of future life episodes.

Mind, Vol. 126 . 501 . January 2017 **© Gerrans and Kennett 2016**
doi:10.1093/mind/fzv206 Advance Access publication 25 October 2016

Our argument drew an analogy between patients with anterograde amnesia and ventromedial patients in respect of their loss of this capacity for what we called diachronic agency. While the deficits are different (hippocampal patients have lost access to the episodic database, while ventromedial patients are unable to integrate affective responses with other cognitive processes) one crucial effect is the same. They cannot effectively project themselves in time. Amnesic patients cannot represent the future episodically, and ventromedial patients cannot imaginatively inhabit their futures.

Ventromedial patients can of course make judgements about future cases using declarative information, but such judgements are not fully agential ones because they are not made by 'feeling the future' (Gerrans and Sander 2014). Thus, although they do have preferences and judgments about moral situations, those judgements are applications of principles or expressions of affectively scaffolded preferences rather than expressions of their capacity for moral agency.

One way to dispute this account would be to argue that diachronic agency is not necessary for moral judgement. Another would be to accept that agency is necessary for moral judgement but to dispute the idea that ventromedial patients lose a capacity for diachronic agency in virtue of their affective deficits. Chris Zarpentine's critique (2016) takes the latter approach. He writes:

> ... vmPFC patients are not impaired with respect to mental time travel and their affective capacities are not completely compromised. As a result, they seem to meet the conditions that G&K claim are necessary for moral agency ... (p. 246)

In effect Zarpentine works his way back from denying (iv) to denying (ii). He argues that the affective deficits of vmPFC patients do not destroy their capacity for mental time travel. Hence their capacity for moral agency is also retained.

He uses the felicitous term 'dynamic evaluation' to describe the kind of affectively-scaffolded evaluative processes which have been the focus of so much research on ventromedial patients. Dynamic evaluation is an 'an ongoing, dynamic assessment of relative value' which depends on the ability to integrate affective responses with representations of relevant information about the future. In this respect all sides are agreed. Loss of the ability to integrate affective responses with representations of relevant information impairs decision-making both *in situ* and for the future.

 © Gerrans and Kennett 2016

The core of Zarpentine's argument is that while ventromedial patients' affective deficits impair their dynamic evaluation, they do not impair their capacity for mental time travel:

> …vmPFC patients are not impaired with respect to mental time travel and their affective capacities are not completely compromised. As a result, they seem to meet the conditions that G&K claim are necessary for moral agency … (p. 246)

The idea is that mental time travel and affective response dissociate. Mental time travel *simpliciter* consists in the *simulation of perceptual and sensory aspects of episodes of future experience*, for example by imagining the view from a mountaintop, or the taste of taro. Dynamic evaluation consists in the association of affective responses with those simulations in order to aid decision-making. And it is the latter capacity that is impaired in vmPFC patients.

Zarpentine's view draws some support from work by Gilboa (2004) and others who distinguish episodic and autobiographical memory. There is a difference between picturing the view from a mountaintop, and picturing the scene while simultaneously evoking the feelings triggered by that view. The latter is what Gilboa calls 'autobiographical' episodic memory, in which we assimilate the experience to ourselves. The important point for decision-making seems to be that the former type of representation, while it provides information, does so in a motivationally neutral way. Episodic representations unaccompanied by affective response are like declarative representations such as words or symbols. They need to be supplemented by affective feeling to incline us to action.

What seems to distinguish episodic from autobiographical representation is the activity of the vmPFC. Gilboa puts it this way:

> Ventromedial PFC is involved in: (i) processing of self-related information; (ii) decision-making under conditions of uncertainty when multiple possible answers are available and (iii) memory control processes providing a 'feeling of rightness'. Considering autobiographical memory, these processes may converge to provide a system that monitors the veracity of retrieved memories. It does so by establishing a *self-related* retrieval template that sets up the parameters against which retrieved memories are evaluated based on (or which give rise to) a general intuitive 'feeling of rightness'. Thus, the monitoring of the veracity and perhaps cohesiveness of autobiographical memories is primarily mediated by the ventromedial PFC. (Gilboa 2004, p. 1346, emphasis added)

 © Gerrans and Kennett 2016

Very crudely, we might say that the vmPFC transforms a purely episodic representation into an autobiographical one by playing its role as a hub in the integration of affective and motivational processing. And this is why vmPFC damage impairs dynamic evaluation.

It does not seem to us to be a stretch, given what we know about the role of the vmPFC in coordinating motivational and affective processing with other cognitive processes required for personal decision-making, to say that it is precisely the integration of affective information which provides the feeling of 'mineness' which transforms an episodic representation into an autobiographical representation. One way of putting this might be that the role of the vmPFC is to make episodic or semantic representations, which would otherwise be motivationally inert, motivationally charged. This was the idea behind the somatic marker hypothesis, which has driven so much work in the field over three decades.

In effect, we argued that vmPFC patients whose capacity for dynamic evaluation was impaired thereby sustained impairments to mental time travel. Zarpentine in contrast argues that:

> … vmPFC patients are not impaired with respect to mental time travel, per se. Rather, their problem with dynamic evaluation results from dysfunction of the mechanisms that generate affective responses. Indeed, they appear to be equally impaired in their ability to engage in dynamic evaluation *in situ* or in response to mental time travel. And this explains their problems with emotional regulation and decision-making. (p. 243, Italics in original)

This view is quite consistent with Saver and Damasio's (1991) report of their vmPFC patient who, as Zarpentine puts it,

> … can imagine future possibilities; he showed normal performance on tasks designed to measure the ability to think of alternative options in a given situation, reflect on the future consequences of actions, engage in means-end reasoning, and foresee the outcome of social situations …(p. 244)

Such a patient is using a mixture of declarative and episodic information *purged of affective and motivational information* to form beliefs about the future. Saver and Damasio do not explicitly credit EVR with preserved episodic memory but even if we assume such abilities are intact EVR could only ever be (and may be) an *episodic time traveller* (or minimal time traveller). He is not an *autobiographical time traveller in the sense described by Gilboa, which invokes the idea of affectively enriched episodic representation used prospectively*. It is very hard for such people, even if they can form beliefs about the future, to project

 © Gerrans and Kennett 2016

them*selves* into future scenarios, in virtue of the loss of affect. Such patients lose, not just their former pattern of affective response, *but a part of themselves necessary for cognitive control and decision-making.* This is why EVR has problems including conduct disorder and affective flattening for personally significant decisions. In effect we argued that vmPFC patients whose capacity for dynamic evaluation was impaired thereby sustained impairments to their capacity for mental time travel.

We can distinguish between minimal episodic mental time travel and affectively enriched autobiographical mental time travel. It is common ground between us and Zarpentine that the latter is impaired by ventromedial damage. Both we and Zarpentine think that ventromedial patients do not 'feel the future'. We agree with Zarpentine that episodic representations purged of affective associations would not allow affective sampling of simulated futures, with consequences not only for decision-making but diachronic agency/dynamic evaluation. The problem is not just that someone who can't feel the future will make bad decisions, but that even if they make good ones using affectless forms of imagination and reasoning, there is no reason to think that they will act consistently with their intention when the projected decision point arrives.

When Tulving (1985) and others coined the term 'mental time travel' the concept of episodic memory had not fractionated: it was used to contrast with declarative memory for facts. Tulving's idea of autonoetic memory suggests he had a concept of affectively enriched autobiographical memory in mind. Such a use would fit the idea that in mental time travel we project ours*elves* in time in order to do some affective sampling to aid decision-making.

However we agree that if the term 'mental time travel' is restricted to the simulation of purely episodic representation in decision-making then it is possible for vmPFC patients' capacities for mental time travel to be preserved. They can imagine what the future would look like even if they cannot feel it. But we think that Zarpentine would agree that what matters for diachronic agency/dynamic evaluation is the ability to feel the future. Indeed our main point was to argue that if people lose this ability they have less of a self to project and hence are compromised as agents. We can also agree with Zarpentine that the extent of that loss and its consequences for moral agency can be matters of degree.

 © Gerrans and Kennett 2016

2. Affective response and learning mechanisms

It is against this background that we make some remarks about the nature of affect itself. Zarpentine criticises us for relying on an impoverished notion of affect leading us to underestimate the capacities for agency of creatures like infants who lack developed capacities for explicit reasoning and mental time travel but can nonetheless care about others, appreciate the significance of actions and use those feelings to regulate their behaviour. His point is that in such cases affective responses, precisely because they are sculpted in an intersubjective developmental process, can provide the basis for a kind of moral agency. The same might be true of patients with memory impairments but intact affective capacities. (Indeed this was our example of an amnesic patient with preserved SCR response.) In such cases preserved affective feelings are more than just reflexive prompts to action because 'such responses actually derive from sophisticated learning mechanisms that are modified over the course of an individual's development and are particularly sensitive to the kind of training that occurs in moral education'.

We agree and are also happy with the idea that agency can come in degrees. However, we sound a note of caution: 'sophisticated learning mechanisms' can install affective responses but once installed they remain reflexive and can function as conditioned responses. This is the plight of Alzheimer's patients: a lifetime of emotional learning reduced to the triggering of feelings toward familiar people. They continue to care but their learning, decision-making and agency are compromised. Things are perhaps better for children in the preschool years, before the onset of a capacity for mental time travel. Their affective responses are indeed being trained up and integrated with their rapidly developing cognitive capacities so they are on an upward trend with respect to their capacity for agency. But surely they are not fully developed until the late primary years.

Our target was the idea, very familiar in the work of neurosentimentalists such as Jonathan Haidt, and in some moods Jesse Prinz, that the cognitive nature of moral agency could be reduced to the reflexive association of affective responses with perceptual or declarative representations—for example, associating a disgust response with a picture of a member of another race to produce a negative moral evaluation. In such cases affective responses are treated as prompts to decision which are encapsulated and inflexible. Their role is to shape

 © Gerrans and Kennett 2016

preferences in ways which are largely opaque to the subject and unmodifiable from the top down.

Antonio Damasio's (1994) breakthrough interpretation of the cognitive nature of the vmPFC was that one of its important roles is integrating emotionally inflected feelings with cognitive processes necessary for planning and decision-making. He went on to describe the role of such integrative processes in the production of an autobiographical self: a representation of the organism as a persisting entity, whose present is emotionally inflected by the past, able to represent itself as continuing into the future. Our aim was to emphasise that the integration of affective responses for agentive purposes actually depends on the ability to feel the significance of actions for one's past and future selves and integrate such feelings with the full range of executive functions. We think that Zarpentine would agree that affect has a more significant role than the determining of a preference ranking for actions by inflexible processes.

However we do differ with Zarpentine on some important points about the nature of affect, which matter to the characterisation of the abilities of what he calls 'marginal agents', such as the Alzheimer's patient and the infant. It may be true that they have intact capacities for care and affective response but it is equally true that they cannot use those capacities for deliberative control which we would say is a prerequisite for agency.

Affective feelings are interoceptive representations (of body state) that carry information about how the subject is faring in the world. Not every representation of body state is an affective representation; much homeostatic representation and interoception is simply involved in monitoring and regulation. However the mind has evolved specialised systems for representing the significance of changes in body state for organismic goals represented at different levels of explicitness and complexity. Thus awareness of raised heartbeat can be simply a matter of ongoing bodily monitoring or, depending on context, it can be part of the feeling of an emotional response to a perceived or imagined situation.

The mind has evolved systems that allow the integration of these emotional feelings with cognition at different levels of explicitness and complexity. Emotional feelings can be integrated with reflexive cognition, which is one aspect of conditioning, or reflective cognition. Crudely put, the amygdala is involved in affective processing for reflexive cognition. It coordinates the rapid online processing of information relevant to appraising the significance ('subjective relevance' is the term

 © Gerrans and Kennett 2016

of art used by appraisal theorists of emotion) of current events. Amygdala based processing can be surprisingly sophisticated, responding to subtle contextual contingencies. Essentially however it is the hub of an online synchronic system since it is driven by perception.

The vmPFC evolved as part of a system of emotional regulation, reappraisal, learning and decision-making, which requires re-evaluating and overriding automatically generated responses. Thus it is essential for reversal learning and extinction of conditioned responses, 'teaching' the amygdala new patterns of response. In humans its regulatory functions have expanded to augment our capacities for higher-level cognitive control by representing the self-relevance of information for reflective processes. Thus the vmPFC essentially recapitulates the role of the amygdala for reflective cognition. The capacities for higher level cognition, reflection and control are, we argue, essential to moral agency and are also the capacities that empirically minded rationalists focus upon.

It is standard to think of the amygdala as a hub of reflexive emotional processing and the vmPFC as a hub of higher level emotional processing in virtue of its participation in so many networks involved in self-referential processes. Our point was that if the vmPFC is damaged then the subject has to rely on the integrity of her lower level hub to control her behaviour. She can no longer dial down reflexive responses and access reflective forms of emotional regulation. And this is the standard explanation of the odd pattern of behaviour of some vmPFC patients. They have flattened responses when thinking about past and future but often heightened, impulsive and disinhibited responses to current events. We think that such people have diminished capacities for agency. Zarpentine writes:

> Rejecting G&K's view of emotion as akin to reflexes and endorsing the account of affective mechanisms as sophisticated learning mechanisms shaped over the course of development can both explain how emotions come to have the requisite structure and how these affective mechanisms can play a role in the constitution of an agent that persists through time (p. 249).

We do not think that emotions are akin to reflexes. *Some* affective processes are automatically integrated with reflexive processing. These are the type emphasised by somatic marker theories and neurosentimentalists such as Haidt. Our point was that agency requires more than that. We think that a capacity for diachronic agency is essential to moral deliberation, but that this capacity is obscured by two features

 © Gerrans and Kennett 2016

of neurosentimentalist accounts. The first is the emphasis on synchronic judgments in hypothetical scenarios rather than protracted, socially embedded, first personal deliberation. The second is the all or nothing contrast between theoretical rationality and affective processing. These contrasts make it seem as if moral decisions can only depend on intuitive processes *or* abstract reasoning. If we focus on the full range of processes involved in moral decision-making we see that the quick affective flashes resulting in intuitive judgments of approval and disapproval that are picked up in the moral judgment research relied upon by neurosentimentalists and cited by Zarpentine, are just one input to moral decision-making.

If affective processes are seen as more flexible, cognitively penetrable and developmentally plastic then they have an essential role to play in constituting agents and selves. In order for them to function that way they have to be integrated with the processes which enable mental time travel. And this is one crucial function of the vmPFC.

Zarpentine concludes that:

> ... given the diachronic aspects of affective mechanisms, neurosentimentalism has the resources to offer a robust account of both how affective mechanisms are partially constitutive of an agent that persists through time and how moral judgements that issue from these mechanisms can be expressions of such agency (p. 251).

We agree. But so, for that matter, does a rationalism which is sensitive to the empirical evidence. As we have suggested elsewhere (Gerrans and Kennett 2006) that if emotions are cognitively penetrable then the contrast between rationalism and sentimentalism diminishes or disappears. And the evidence suggests that vmPFC damage does impair the essential role of affective mechanisms, which is to say it damages the capacity for emotion regulation, revision and control.

References

Damasio, Antonio 1994: *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putnam.

Gerrans, Philip and Jeanette Kennett 2006: 'Introduction'. *Philosophical Explorations*, 9, pp. 3–12.

——2010: 'Neurosentimentalism and Moral Agency'. *Mind* 119, pp. 585–614.

 © Gerrans and Kennett 2016

Gerrans, Philip and David Sander 2014: 'Feeling the Future: Prospects for an Implicit Theory of Prospection'. *Biology and Philosophy*, 29, pp. 699–710.

Gilboa, Asaf 2004: 'Autobiographical and Episodic Memory—One and the Same? Evidence from Prefrontal Activation in Neuroimaging Studies'. *Neuropsychologia*, 42, pp. 1336–49.

Julian C. Motzkin, Carissa L. Philippi, Richard C. Wolf, Mustafa K. Baskaya, and Michael Koenigs, 2014: 'Ventromedial Prefrontal Cortex Lesions Alter Neural and Physiological Correlates of Anticipation'. *The Journal of Neuroscience*, 34(31), pp.10430–7.

Saver, Jeffrey L. and Antonio R. Damasio 1991: 'Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy due to Ventromedial Frontal Damage'. *Neuropsychologia*, 29, pp. 1241–9.

Tulving, Endel 1985: 'Memory and Consciousness'. *Canadian Psychology*, 26, pp. 1–12.

Zarpentine, Chris 2016: Moral Judgement, Agency and Affect: A Response to Gerrans and Kennett. *Mind*, 126, pp. 233–57.

© Gerrans and Kennett 2016

Book Reviews

One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness, by Graham Priest. Oxford: Oxford University Press, 2015. Pp. xxviii + 252.

How is it that a multiplicity of distinct objects can co-operate to form one object? What accounts for the unity of a complex object, for its being a single thing and no mere disparate plurality of its various parts? In this bracingly original treatise, Graham Priest sets out to answer these questions and others intimately connected with what it is to be metaphysically *one*.

Reflection on these questions, Priest argues, discloses an aporia: it appears that whatever accounts for a complex object's unity – that object's *gluon*, in Priest's nomenclature – must itself be, and yet cannot be, an object. (This paradox featured briefly in Priest's earlier *Beyond the Limits of Thought* (Oxford: OUP, 2002, p. 193), where the term 'gluon' was adopted with apologies to particle physics.) Deploying the view for which he is best known, dialetheism, Priest advances an explanation of metaphysical unity which embraces each of these contradictory appearances as veridical: the gluon of a complex unity both is and is not an object. Priest proceeds to argue that his dialetheic explanation of unity – *gluon theory*, as he dubs it – has fruitful applications for a wide range of problems in metaphysics, philosophy of language, the history of both Western and Eastern philosophy, and even ethics.

We'll come to the tenets of gluon theory momentarily. First, consider the alleged aporia. Suppose that a complex object, *O*, comprises the parts, *a*, *b*, *c*, and *d*. According to Priest, something – *O*'s gluon, *g* – must explain *O*'s being a unified whole and no mere congeries of *a*, *b*, *c*, and *d*: something must bind together *a*, *b*, *c*, and *d* in the unitary object, *O*. But then *g* cannot, Priest alleges, simply be another object alongside *a*, *b*, *c*, and *d*; for if it were, its invocation would leave us with just another plurality of objects – *a*, *b*, *c*, *d*, and *g* – and not the unity *O*:

> If [the gluon] were [an object], the collection of parts *plus* the gluon would constitute a plurality just as much as the original. So the problem of binding would not be solved. (p. 9, emphasis in original)

© Mind Association 2016

Just as there had to be something which bound together *a*, *b*, *c*, and *d*, similarly, Priest reasons, if *g* is an object,

> … there must be something which binds *g* and *a*, *b*, *c*, *d* together, a hyper-gluon, *g′*. There must, then, be something which binds *g′*, *g* and *a*, *b*, *c*, *d* together, a hyper-hyper-gluon, *g′′*. Obviously we are on an infinite regress. Moreover, it is a vicious one. (p. 11)

Thus, Priest concludes, *g* is not an object. And yet it is! For, on the thin conception of objecthood at stake, our having just singularly referred to it suffices for it to be an object (a consideration, Priest notes, that caused Frege and Heidegger grief in their reflections on the concept *horse* and being, respectively).

Gluon theory boldly bites this aporetic bullet. The key idea is that every complex object (concrete or abstract) is unified by virtue of its having a part – its gluon – that is numerically *identical* to *each* of that object's parts, including its improper part. The object's being *one* is owed to there being some one thing that is the very same thing as the object and each of its parts. The aforementioned regress is halted by gluon theory after its first step: no hyper-gluon is needed to unite *g* and *a*, *b*, *c*, *d*; for *g* is none other than *a*, none other than *b*, and so on. There is no 'metaphysical space' (p. 17) between an object's gluon and each of that object's parts, across which some additional join is needed to secure unity.

'Wait', one naturally protests, 'the phenomenon of which we sought an explanation was *complex* unity: the holding together of *distinct* parts in the whole. If the parts are each identical to some one part, they are all identical, not distinct. This is absurd: it implies, e.g., that Priest's left foot is identical to his right'. Priest's daring response is to deny the transitivity of identity: though a complex object's parts are each identical to its gluon, they are not identical to one another. Moreover (to recount Priest's reasoning rather too briskly (see pp. 26-7)), since a gluon, *g*, is identical to distinct parts, *a*, *b*, which, being distinct, differ in respect of some property; and since *x* and *y* are identical iff *x* has a property just if *y* has it; it follows that *g* differs from itself in respect of some property, and thus is not identical to itself. But *g* is not identical to anything if it is not identical to itself, and since to be an object is to be identical to something, *g* is therefore not an object. By the same token, however, since *g* *is* identical to something (e.g. *a*), it *is* an object. Thus, as the aporia appeared to indicate, gluons indeed both are and are not objects.

Priest proceeds to apply this explanation of unity in treating of a formidable array of conundrums drawn from an uncommonly wide range of traditions and periods in philosophy. He applies gluon theory in developing, inter alia: a theory of universals (Ch. 3) purporting to overcome problems raised in Plato's *Parmenides* (Ch. 7) (briefly: the universal of, e.g., redness is the gluon identical to all and only redness tropes); an account of the Heideggerian preoccupations, being and nothing(ness) (Ch. 4) (briefly: the being of an object is simply its gluon; nothingness is the gluon of the fusion of the empty set); and a structuralist interpretation and defence of the view, central

 © Mind Association 2016

to the Madhyamaka school of Buddhist philosophy, that all things are empty (Ch. 11-13). The book closes with reflections on the ethical implications of the foregoing metaphysics, arguing that the emptiness of all things constitutes the ground of compassion (Ch. 14-15).

The book, it has to be said, is veritably teeming with logico-metaphysical heresies. Several have already been mentioned: some contradictions are true; identity is not transitive; some objects are not self-identical; the empty set has a mereological fusion. There are many others, however: some objects don't exist, but it is possible to perceive them nonetheless (passim, p. 158); parthood is not anti-symmetric (p. 89); set-membership is not well-founded (pp. 191ff); there is a universal set, and it has a fusion (p. 100); there is a mereological fusion that neither overlaps itself, nor is an improper part of itself (pp. 91, 97ff.); identity needn't be necessary or permanent (p. 26); Plato was a dialetheist (Ch. 8); every object is ontologically prior to (ontologically dependent upon) itself (p. 187); there are things that are absolutely ineffable (p. 200).

Despite what this list of Priestly heterodoxies might suggest, however, the author never gives the impression of indulging in iconoclasm for its own sake. The book is certainly radical in its contentions; but, far from being separate assaults on received metaphysical wisdom, Priest's unorthodox views are intimately (and aptly) bound together: accepting one often makes it very natural to accept several others. Moreover, Priest's characteristic logical rigour – his policy of presenting precise formal models to demonstrate the coherence of his claims – offers reassurance that his dissenting views are, at least, sane and sober options.

That said, given how far many of Priest's views depart from the consensus, his defence of these views and discussion of potential objections is sometimes unsatisfyingly brief. His rejection of the substitutivity of identicals, of which the transitivity of identity is a special case, is an example (Ch. 5). Priest appeals here to a number of much-discussed cases involving intentionality, time, modality, fission, and vagueness. However, in each case Priest neglects, or gives only cursory consideration to, important lines of response that are likely to jump out at readers familiar with the relevant debates. The author does warn that he makes no attempt at a comprehensive discussion, and his brevity is, I think, both made inevitable, and compensated for, by the remarkable breadth of the book.

The only substantive point of criticism I wish to make concerns the putative aporia that sets the whole inquiry in motion. Priest fails, in my view, to make a convincing case that there is any aporia concerning the objecthood of gluons. In particular, he provides no good reason to think that gluons are not objects. Return to our complex object, *O*. Why should it trouble us that if *O*'s gluon is an object, then it *plus* *O*'s parts 'constitute a plurality'? Of course they do; since – as I understand Priest's terminology – for some things to constitute a plurality is nothing more than for them to *be some things*. Surface grammar notwithstanding, it is not for them to constitute some *one* thing,

 © Mind Association 2016

called a plurality. Crucially though, their constituting a plurality is perfectly *consistent* with their also constituting some one thing, some unity: that they constitute a mere plurality does not entail that they *merely* constitute a plurality. They may also form a unity, and, for all Priest has said, may do so by virtue of the unifying potency of the gluon. It's not clear, then, that there is any problem with supposing that gluons are objects.

Perhaps, though, the problem really lies with the vicious regress, unmistakeably redolent of Bradley's, which is allegedly ignited by the supposition that gluons are objects. Having posited a gluon, *g*, which binds together *a*, *b*, *c*, and *d*, we must now, says Priest, posit a hyper-gluon, *g*′, which binds together *g* and *a*, *b*, *c*, and *d*; and off we go. However, there is an unwarranted tacit assumption in this part of Priest's argument. We only have to posit the aforementioned hyper-gluon if we think that *g*, and *a*, *b*, *c*, *d* together constitute a unity; and why need we think that? After all, our supposition was only that *a*, *b*, *c*, and *d* constitute a unity.

To illustrate, suppose we hold that what binds together *a*, *b*, *c*, and *d* as a unity is the *fact* that they stand in some relation, *R*; moreover, we deem the fact that *Rabcd* to be an object. We need not think that *a*, *b*, *c*, *d* and the fact that *Rabcd* constitute some unity (indeed, if *a*, *b*, *c*, and *d* are not themselves facts, common sense favours refusing to do so); so we need not recognize some gluon that binds them together. On this view, then, the regress can halt after the first stage. I'm not claiming that the view is an ultimately satisfactory account of unity, only that it is altogether unclear that the objecthood of a gluon would compel us to recognize an additional gluon binding it together with the parts of the original unity.

I don't think this effaces the interest of gluon theory. The theory is an admirably elegant account of metaphysical unity that deserves serious consideration. I only suggest that the argument purporting to force us into a dialetheic metaphysics of objects-that-are-not-objects is weaker than Priest thinks.

Nevertheless, the breadth and boldness of this work, as well as its technical rigour and historical sensitivity, are very much to be admired. I doubt that this latest treatise will independently win many converts to dialetheism, but the book is an important addition to the impressive cumulative case for true contradictions which Priest has been assembling over several works. It is certainly deserving of the attention of those working in metaphysics and related fields.

Corpus Christi College
Oxford
OX1 4JF
michael.price@philosophy.ox.ac.uk
MICHAEL PRICE

doi:10.1093/mind/fzw043 Advance Access publication 23 December 2016

 © Mind Association 2016

The Laws of Belief: Ranking Theory & its Philosophical Applications, by Wolfgang Spohn. New York: Oxford University Press, 2012. Pp. xv + 598.

Probabilists model partial belief by assigning it a degree between 0 and 1; the higher the degree, the more credible the proposition. In contrast, ranking theorists assign numbers or ranks to propositions so that the higher the number, the more the proposition is *dis*believed. In *The Laws of Belief*, Wolfgang Spohn develops ranking theory in detail and applies it to a host of philosophical problems. The book has an extraordinary sweep, taking in many areas of metaphysics, epistemology, and philosophy of science and even the philosophy of language. Long in the writing, it synthesises thirty years' work on ranking theory.

A ranking function κ is a function from possible worlds to the natural numbers 0, 1, 2, 3, ... and ∞ (infinity). Let W be a set of possible worlds and A an algebra of propositions over W, a proposition being a subset of W. The ranking function κ is extended from worlds to propositions by setting $\kappa(A) = \min\{\kappa(w): w \text{ is in } A\}$, from which it follows that $\kappa(A_1 \vee A_1) = \min\{\kappa(A_1), \kappa(A_2)\}$. This gives a grading of disbelief: $\kappa(A) = 0$ means that the proposition A is not disbelieved at all (A may be believed or neither believed nor disbelieved), $\kappa(A) = 1$ means that A is disbelieved to the least degree, $\kappa(A) = 2$ that A is disbelieved to the second least degree, ..., and $\kappa(A) = \infty$ that A is utterly disbelieved. Contradictions have rank (i.e. κ-value) ∞ and are therefore utterly disbelieved, whereas tautologies have rank 0. The subject believes A if and only if $\kappa(\neg A) > 0$; equivalently, if and only if she does not disbelieve A and also disbelieves $\neg A$ to some extent.

How are ranks affected by new evidence? Suppose your doxastic state at a particular time is represented by the ranking function κ. You then receive evidence E with firmness n. How should you (rationally) revise your ranking function? Spohn's answer is: if E contains w then w's new rank is $\kappa(w) - \kappa(E)$, the conditional rank of w given E; and if E doesn't contain w, then w's new rank is $\kappa(w) - \kappa(\neg E) + n$. This principle can be thought of as a ranking-theoretic analogue of Jeffrey conditionalisation. As Spohn shows, this account generalises the well-known AGM revision rules.

Spohn applies ranking theory to the notions of justification, the a priori, laws of nature, confirmation, causation, and many other philosophical topics besides. A simple application that illustrates some of ranking theory's basic ideas is the lottery paradox. Consider a lottery with N tickets, in which A_i is the hypothesis that ticket i will win for $i = 1$ to N. Each of the following claims is highly plausible: (i) it is rational to believe $\neg A_i$ for each i; (ii) if it's rational to believe X and it's rational to believe Y then it's rational to believe $X \wedge Y$;

 © Mind Association 2016

and (iii) it's not rational to believe $\neg A_1 \wedge \ldots \wedge \neg A_N$, since not all tickets will lose. Clearly, however, (i)–(iii) are inconsistent; hence the paradox. From the perspective of ranking theory, the key observation is that:

$$0 = \kappa(A_1 \vee \ldots \vee A_N) = min\{\kappa(A_i) : 1 \leq i \leq N\}$$

The first identity follows from the fact that $A_1 \vee \ldots \vee A_N$ is certain in the given scenario, and the second from the axiom for finite disjunction. Thus $\kappa(A_i) = 0$ for some i, so $\kappa(\neg\neg A_i) = 0$, i.e. not $\kappa(\neg\neg A_i) > 0$ for this i. Since a hypothesis is believed iff its negation has non-zero rank, it follows that $\neg A_i$ is *not* believed for some i. In other words, according to ranking theory, one cannot rationally believe of each of a finite lottery's tickets that it won't win (assuming one believes the disjunctive hypothesis that one of them will win). Ranking theory thus resolves the paradox by rejecting (i).

Above, I presented the codomain of the ranking function as the natural numbers with infinity adjoined. This is not essential: other possible codomains include the reals or ordinals with infinity adjoined. Flexibility over the choice of codomain enhances ranking theory's range of applications, as long as the choice of codomain can be justified by an appropriate measurement theory. This brings us to an important question: where do ranks come from? Probabilists have a natural way to connect degrees of belief with behavioural manifestations (betting behaviour, etc.), but what about ranking theorists? What does it mean to say that the rank of hypothesis A is 3, whereas that of B is 5, or that evidence E is received with firmness 8? Even if one takes the line that subjective ranks are primitive and irreducible to behaviour, they must somehow be connected to empirically observable phenomena.

The task of providing a measurement theory of ranks falls to chapter 8. Spohn offers three accounts, the third of which is his official view. The discussion is too technical to even begin to do justice to it here. Suffice it to say that the key idea is to link the notion 'A given B is at most as disbelieved as C given D' to facts about iterated belief contraction. (A belief contraction occurs when you go from believing some hypothesis to suspending judgment about it.) Given some conditions on the four-place relation 'A given B is at most as disbelieved as C given D', standard measurement-theoretic results guarantee the existence of a rank function from the algebra of propositions to the real numbers with infinity adjoined. However, as Spohn points out (p. 178), the method gives us no way to compare ranks across different subjects. Ranks are therefore like utilities, which are not interpersonally comparable, but unlike probabilities, which are absolute and hence interpersonally comparable.

Spohn's ranking theory has several antecedents. Perhaps the most important is the work of the twentieth-century English economist G.L.S. Shackle. Shackle developed his theory of 'potential surprise' most fully on pages 79 to 113 of his *Decision, Time and Order*, drawing on material already to be found on pages 128 to 133 of *Expectation in Economics* and on briefer remarks in

 © Mind Association 2016

chapter 2 of *Time in Economics*. What Spohn calls a rank is what Shackle called a degree of potential surprise, the lowest degree being 0 and the absolute maximum of potential surprise being assigned to hypotheses that are absolutely disbelieved, i.e. hypotheses with infinite rank in Spohn's terminology. Many other ideas are common to the two accounts, including the condition that at least one member of an exhaustive set of rival hypotheses has lowest rank (in Spohn's terminology) or carries zero potential surprise (in Shackle's). The main foundational difference that distinguishes ranking theory from Shackle's axiomatic theory of potential surprise is the account of conditionalization. As Spohn observes, 'whatever is novel in this book ultimately derives from the explanation of conditional rank' (p. 78). It is perhaps surprising that Spohn first conceived of ranking theory in 1982 (p. xii) without apparent awareness of Shackle's work, since the latter had some currency in the 1950s and 1960s and had even been discussed by philosophers of rational belief revision such as Isaac Levi. (See pages 135 to 138 of the latter's influential 1967 book *Gambling with Truth*, and the preceding account of *d*-measures therein). In any case, Spohn generously acknowledges Shackle's work as 'remarkably early and clear-sighted' (p. 229), though to add that Shackle did not really go beyond basics in his development of the key idea (p. 229) may be to undervalue his contributions.

Specialists in rational belief revision theory will of course want to get hold of this book and pore over it—as the cliché goes, it is required reading. As for non-specialists, the author makes an effort to summarise the state of the art in rational belief revision theory and in the various fields to which he applies ranking theory: see the introductory sections 12.2 for lawlikeness, 13.1 for ceteris paribus laws, 13.4 for dispositions, 14.2–4 for causation, 15.1 for objectivisation and 16.1 for justification. Despite this, honesty bids me to say that *The Laws of Belief* is really for formal epistemologists only. One prohibitive feature is the book's length: 555 single-spaced small-font pages. Another is the presentation of the material, which leaves much to be desired. The author is at times expansive when brevity would have been advisable, and brief when we should have liked him to be more expansive. The prose is not always easy to read, and the mathematics, though not especially difficult, is presented with little by way of informal explanation or motivation. Another difficulty stems from the author's tendency to gesture in the direction of the literature. An example is his response to the first objection to Jeffrey conditionalization, roughly that the primitive assignment Q of probabilities to the evidence ought to be the result of an interaction between the prior probabilities and the evidential process. Spohn's first response takes up a single sentence:

> But the objection can be met: there is a way of characterizing the observational import and its interaction with the prior probabilities that yields for any prior P a posterior measure Q on E (cf. Field 1978; see also the discussion of Garber 1980 and Jeffrey 1983, pp. 181–183). (pp. 39–40)

 © Mind Association 2016

How this is supposed to work will surely be opaque to anyone who does not know or cannot recall the cited literature. As Spohn acknowledges a little later (p. 41), his discussion here and occasionally elsewhere proceeds more by gesturing to the literature than by self-standing explanation.

One way to draw in non-specialists is by taking early steps in the development of the subject very slowly and persuasively. With an inviting and careful introduction, they might be led further in. We look then to the discussion of the tenets or 'laws' of the whole enterprise, as Spohn articulates them (p. 48):

> (A1) Belief sets are consistent.
> (A2) Belief sets are deductively closed.

Since beliefs are propositions one might express in natural language, a natural question to ask regarding (A1) is: what is it for a set of propositions to be consistent? We know what it might be for a set of formal sentences to be consistent relative to a deductive system: usually this means that the deductive closure of the set does not contain $p \wedge \neg p$ (or some equivalent condition). But what about, say, English? What are the rules of inference and axioms of English—what, if any, is the *logic* of English? Or perhaps the notion of consistency in (A1) should not be understood in a formal sense. Perhaps it is best construed as metaphysical consistency, where a metaphysically consistent set is one all of whose elements are true in some metaphysically possible world. Similar questions arise for (A2).

So there are foundational questions aplenty at the very start of the enterprise. Unfortunately, these are skimmed over quickly, in a series of brief and imperspicuous remarks. Raising a potential worry about (A2), which he will go on to dispel, Spohn writes:

> Still, logical consequence is undecidable (in the technical sense). Hence, (4.2) [A2] assumes a belief set to be objectively, and a fortiori, subjectively undecidable. What should it mean then that a believer is disposed to believe such an undecidable set? (p. 48)

His justification of the two fundamental assumptions is then twofold. First, a theory that assumes (A1) and (A2) is said to be incomparably more fruitful than one that does not. The second justification is that they flow from the identification of the objects of belief with propositions understood extensionally as truth-conditions, so that for example equivalent representations of the same truth condition are always recognised as such (p. 49).

To try to make sense of this, let's look again at the first sentence of the quotation from page 48. When Spohn says that logical consequence is undecidable, presumably he has in mind the undecidability of first-order logic's consequence relation restricted to finite premiss sets, or the undecidability of a similar logic. How this relates to English remains unclear, unless he has tacitly assumed that first-order logic perfectly models English consequence. And the question broached earlier of what exactly is meant by the 'deductive closure' of a set of natural-language sentences remains unanswered. The

 © Mind Association 2016

notion of logical consequence Spohn relies upon is thus left implicit and, to my mind, unclear. At the bottom of page 49 and the top of page 50, we are told that deductive closure boils down to recognising that $p_1, \ldots, p_n$ are true if and only if $p_1 \wedge \ldots \wedge p_n$ is true. But that involves two assumptions, neither of which is flagged up as such. First, it is to conceive of logical consequence propositionally. Yet the propositional consequence relation, for finite premiss sets, is decidable, contradicting Spohn's earlier claim of undecidability. Second, it is to assume that logical consequence is compact (since premiss sets are assumed to be finite). What underpins each of these assumptions is mysterious.

Returning to the quotation from page 48, Spohn leaps from the undecidability of logical consequence to the undecidability of any belief set. But this is an invalid inference, since some belief sets may be decidable even if the consequence relation is undecidable. For example, several first-order theories are decidable (e.g. Presburger Arithmetic), even though the first-order consequence relation restricted to finite premiss sets is undecidable. In other words, even if we assume that the logic of belief sets is first-order logic, it does not follow that any belief set is undecidable. Furthermore, Spohn does not explain why it would be problematic to model a rational subject as having an undecidable belief set—this is left unexplained. Perhaps a rational subject could have a decision procedure that determines of any proposition that it is a belief of theirs, but fails to determine the doxastic status of some non-beliefs. A rational subject's beliefs about arithmetic might for instance consist of all and only the theorems of first-order Peano Arithmetic, which is a semi-decidable set: one can generate all such theorems effectively. Once more, the reader is left in the dark about the justification for Spohn's acceptance or rejection of foundational claims.

What, finally, is implied by identifying objects of belief with truth-conditions? Spohn seems to think that deductive closure of the belief set falls straight out of an extensional construal of beliefs. Closure is reduced to the subject's recognition that '(i) if the actual possibility is in each of two propositions, then it is also in their intersection, and (ii) if the actual possibility is contained in some proposition, then it is also in any superset' (p. 49). But this justification of (A2) is circular: it justifies the closure of beliefs from the logical closure of beliefs construed extensionally, given set-theoretic definitions. The first assumption is that if each of $x \in A$ and $x \in B$ is believed then so is $x \in A \cap B$, the last proposition being a logical consequence of the first two given the definition of set intersection. The second assumption is that if $x \in A$ is believed and $A \subseteq B$ then $x \in B$ is also believed, which also follows logically from the definition of a superset. It is hard to see this as a justification, rather than a restatement, of logical closure.

Spohn's view also seems to imply that a subject has very many beliefs. If the objects of belief were, as standard, sentences of a countable language or were

 © Mind Association 2016

expressible by such sentences, then a closed belief set would at most be countably infinite. But if, as for Spohn, the objects of belief are classes of possible worlds closed under intersection and superset relations, then a subject will have as many beliefs as the power set of the class of possible worlds. This follows from the natural assumption that the subject has at least one belief the complement of whose content is no smaller than the infinite class of all possible worlds. (For example, let p be the set of worlds in which Queen Elizabeth II wore a yellow dress on her 90^{th} birthday; the set of not-p worlds is no smaller than the (infinite) class of all possible worlds.) Assuming there are infinitely many possible worlds, the belief set is then at least continuum-sized; and if there is an absolute infinity of possible worlds, its size would be greater than any cardinal. The fact that beliefs must form a sigma-algebra may be invoked to prune belief sets, but this would fly in the face of Spohn's avowed justification of closure.

I have focussed on Spohn's exiguous discussion of the basic ideas in the field. I do not intend to suggest that he does not have sensible answers to such simple questions, nor that he hasn't thought about them a great deal. Rather, it is to indicate that the book does not cater for those not already steeped in this area of formal epistemology. (That said, much of the belief revision literature is similarly cavalier about these or kindred foundational assumptions.) Though the book is a missed opportunity in terms of bringing ranking theory to a broader public, let me stress in conclusion that it is an extraordinary synthesis of decades of research and that it represents an important contribution to formal epistemology.

References

Levi, I. 1967, *Gambling with Truth* (London: Routledge & Kegan Paul).
Shackle, G.L.S. 1949, *Expectation in Economics* (Cambridge: CUP).
——1958, *Time in Economics* (Amsterdam: North-Holland).
——1969, *Decision, Time and Order*, 2nd ed. (Cambridge: CUP).

Wadham College, Oxford A. C. PASEAU
alexander.paseau@philosophy.ox.ac.uk
doi:10.1093/mind/fzw047 Advance Access publication 28 September 2016

Kant's Transcendental Deduction, by Henry Allison. Oxford: Oxford University Press, 2015. Pp. Xv + 477.

The purpose of this book, by one of the foremost contemporary Kant scholars, is to advance a simultaneously 'analytic and historical' treatment of the most complex component of Kant's philosophy: the Transcendental Deduction (p. 1).The approach is analytic in that Allison seeks to provide a rigorous

 © Mind Association 2016

assessment of the multiple lines of argument visible in Kant's text. It is historical in that this is embedded within a detailed exegetical framework: Allison considers, often line by line, a huge range of sources, beginning with the Prize Essay before moving through the A Deduction and the *Prolegomena* to the B Deduction. This interweaving of the textual and the conceptual makes the book immensely valuable: by closely tracking the development of Kant's thought, from its emergence from rationalism through to the modifications implemented over the two editions of KrV, Allison provides a rich picture of the tensions and pressures which shape the Critical system. My aim is to provide an outline of Allison's analysis, and to identify some challenges which his approach faces.

Before I proceed, some remarks on scope. The book is a long one – over four hundred and fifty pages. I cannot do justice here either to its richness or to the field in which it is operating. There will therefore be three major limitations on what follows. First, I will focus on the philosophical side of Allison's reading, rather than the textual details: many of these hang on individual lines which could, and sometimes have, merited a monograph each (for example, the infamous B160-1n). Second, I concentrate on Allison's treatment of the B Deduction since he, like the vast majority of commentators, regards this as the best expression of Kant's argument. (Heidegger is the most influential dissenter.) Third, I am not going to discuss transcendental idealism. Allison's view is that the Deduction and Kant's idealism are 'inseparable' (pp. 453, 286). As he acknowledges, this would for many suffice to constitute a reductio of the former – since Strawson, much of Anglophone Kant scholarship in particular has been predicated on disentangling the two. How one sees this issue obviously depends on what exactly transcendental idealism amounts to, something about which Allison has famous views, but I cannot address these here.

I begin with some background on Allison's reading of the pre-Critical Kant, and on the structure of the B Deduction itself. I will then highlight four aspects of Allison's discussion which seem both particularly interesting and potentially problematic.

The opening third of Allison's text offers a detailed and helpful analysis of Kant's intellectual development. Allison stresses three aspects of this trajectory, all of which seem rightly central. The first is Kant's struggle with the question of method. As Allison eloquently charts, the Kant of the early 1760s frames this in terms of a contrast between 'analytic' and 'synthetic' approaches, labels which have a very different meaning from that which they have in KrV. The analytic method is effectively a regressive one, moving from the observation of particular events to the rules or laws underlying them: the exemplar which Kant has in mind is, of course, Newton. The synthetic method, in contrast, exemplified by work in mathematics, 'begins with arbitrarily chosen definitions from which it deduces consequences by logical means' (p. 12). Allison brings

 © Mind Association 2016

out with particular clarity the complex dynamics engendered by the combination of Kant's rationalist heritage with his insistence that '[t]he true method of metaphysics is basically the same as that introduced by Newton' (U: 286). How, for example, should we understand Kant's views on the observed event, or 'immediate and self-evident consciousness' which serves as the equivalent of Newton's empirical observations, i.e. as the point from which philosophy should regress (p. 12)?

The second, and closely connected, issue emphasised by Allison is the relationship between logic and the broader idea of grounding: as he notes, the rationalist tradition effectively analyses the latter in terms of the former, and Kant initially follows this, albeit with some variation as to the nature of the logical principles involved (p. 15). The evolution of Kant's work in this regard has been widely discussed, but Allison's informative discussion provides a helpful framework. This is particularly the case with respect to the key essay on negative magnitudes, which constitutes a crucial step in the process of disengaging from rationalism by forcing the issue of non-logical grounding and opposition (NG: 171; 203). As Allison notes, the issue here is tightly interwoven with the question of method: in many ways this essay represents 'a last-ditch attempt to deal with the concept of a real ground within the framework of the analytic model of cognition' (p. 22).

The third strand of the pre-Critical analysis concerns Kant's growing awareness of the need for a fundamental split between two kinds of content: intuition and concepts. Kant famously invokes this in KrV to distance himself from both rationalism and empiricism conceived of as reductionist projects operating in favour of one class or the other (KrV: A271/B327). Allison argues that the key developmental moment here is the widely discussed Reflexion 5037 (Refl. 18:69) which identifies 1769 as having given 'great light'. For Allison, what Kant saw that year was precisely the 'discursivity thesis', or the irreducibility of the two classes of content and the central need for some story about their cooperation (p. 47). In contrast, Allison suggests, convincingly I feel, that it is wrong to use that note's reference to 'an illusion of the understanding' as the basis for attributing a 'clear recognition of the antinomial problem' to Kant at that stage (p. 49).

Bringing these points together, we arrive at some of the basic parameters of the Transcendental Analytic: an acceptance of multiple forms of content and the need for a story as to their interaction; an awareness that real grounds need a non-rationalist explanation; and an underlying recognition that these issues can only be treated in the context of a radically new methodology. Some of Allison's historical claims will doubtless attract vigorous responses from scholars concerned with the relevant sub-issues: for example, he is harsh on those who cast Tetens as 'model' for Kant's picture of the imagination (p. 157). But I found the broad story he tells about Kant's development insightful. This holds both at the granular level (for example, the status of 'symbolic cognition' in the *Dissertation* at pp. 63-5 or p. 297 on '*Bewußtsein*

 © Mind Association 2016

überhaupt'), and with respect to the larger trajectory of Kant's thought (for example, Allison offers a neat alternative to Carl's reading of apperception in the *Duisburg Nachlass*, playing down the links to rational psychology, pp. 121-30).

This brings me to the Deduction itself: Allison goes through both editions, at times almost line by line. As stated, though, I will focus on the B version. One immediate virtue of his approach is that he swiftly sets aside readings which have little merit and yet which often clog up the debate: for example, that Kant simply equates the categories with the logical forms of judgment, or that KrV's talk of apperception somehow rests on the kind of Cartesian self which the Paralogisms attack shortly after (pp. 171, 246). The basic structure of the argument for him follows Henrich's now standard 'two steps in one proof' model (p. 327). The 'first step argues that the categories are rules for the *thought* of an object of sensible intuition in general' (original emphasis, pp. 328-9): Kant is thus able to 'conduct this analysis on the minimalist assumption that some data are given, without considering how they are given, e.g., as in space and time' (p. 7). The second step then switches the focus to the specific forms of intuition identified in the Aesthetic. There are two crucial reasons for this. On the one hand, there is a danger that those forms might be in tension with the demands of the understanding: indeed, 'the overall goal of the Deduction is to remove the worry that what is given in sensible intuition might not conform to the categories' (p. 328). On the other hand, attention to those forms allows us to identify certain distinctive modes of intuitive unity which are not only compatible with category deployment, but actually require it.

So, for example:

> the necessity of representing entities and occurrences as occupying determinate positions in a single time, unlike unifying them on the basis of the categories alone, is not imposed upon them by the nature of the understanding, but by the nature of time. (p. 388)

The Aesthetic thus functions as a second source of pressure mandating category use, albeit in some type of pre-discursive mode (p. 329). Allison's overall view of the B Deduction seems to me a plausible one; it is not that distant from that defended by commentators such as Longuenesse (for example, Longuenesse 1998, p. 243), with many of the differences hanging on the details of texts such as B160-1n. At an exegetical level, it allows him to do a good job of ironing out the prima facie non-sequiturs any Kant interpreter is faced with. Allison sees many of these as stemming from ambiguities in the notion of objecthood (e.g. pp. 353-4), and this seems exactly right to me: the Deduction is effectively an extended argument that seeks to move between various non-equivalent notions of objecthood by appealing to their relations to other unities, in particular the spatio-temporal unity of the Aesthetic, and the unities of consciousness and of judgment.

 © Mind Association 2016

It is now time to offer some assessment of Allison's arguments themselves. As he notes, Kant gives multiple formulations of the Deduction's goals, not all of which are obviously equivalent. In what follows I will assume that its central aim is to 'demonstrate the necessary conformity of appearances to the categories, which would establish their a priori objective validity for all appearances and only for appearances' (p. 434; similarly p. 328).

The first aspect of Allison's argument that I wish to discuss concerns the role of empirical cognition. Crudely, Kantian concepts serve several related functions. Most obviously, they are pieces of generic content, representations of the 'characteristic marks' of things, through which we order intuitions in line with a series of inferential or proto-inferential 'rules' (Refl.16:300; KrV:A106; A126). So, for example, an ox can have a 'clear representation', understood phenomenologically, of some property of its stall, such as having a door (SvF: 59). This is the basis for both differential reaction (the ox would behave differently in a stall with no door), and for association (the ox becomes anxious or excited depending on past experiences with doors). What distinguishes concepts:

> does not consist in the fact that that which is a characteristic mark of the thing is clearly represented, but rather in the fact that it is recognized [*erkannt*] as a characteristic mark of the thing. (SvF: 59)

This recognition entails both an awareness of the mark as an iterable property and an awareness of a set of implications as grounded in it: for example, insofar as I conceptualise something as a door, I must either attribute further properties to it, or revisit the initial attribution (KrV: A106). It is in this sense that the *Logic* identifies marks as both 'in the thing [*Ding*]' and as a 'partial representation…considered as the ground of cognition' (Log.:58). In virtue of these features, mark recognition imposes a normative order on experience, preventing it from being 'haphazard' (KrV: A104). More broadly, conceptuality implies at least a tacit awareness of the distinction between inferences putatively grounded in the 'thing' and other ways of combining representations, such as association, which are not so grounded. In this sense 'judgment is nothing other than the way to bring given cognitions to the objective unity of apperception', allowing me 'to say that the two representations are combined in the object' (KrV: B141-2). Finally, to conceptualise is to be aware, again if only tacitly, that insofar as an inference is presumptively objective, as opposed to being merely an artefact of my own psychological history, it should hold for any other observer, 'regardless of any difference in the condition of the subject' (KrV:B141-2). Thus:

> [t]he representation of the manner in which various concepts (as such) belong to a consciousness (in general, not only my own), is judgment. (Refl.16:633).

Or as Allison puts it:

> the basic idea is that in predicating a property or set of properties of some object x, I am implicitly expressing an epistemic ought, claiming that the predication holds

 © Mind Association 2016

> not only for me in my present perceptual situation but for any observer, which makes it objectively valid. (p. 441)

These lines of thought raise familiar textual problems; particularly, with respect to the relation between the *Prolegomena* and the B Deduction. Allison provides a plausible treatment of those issues (pp. 362-70), although readers will doubtless have their own preferred fixes for the many fine-grained exegetical issues here. But what I want to focus on is the argument that Allison offers from these facts about conceptuality to the categories.

> [T]he crucial point is that, as rules, empirical concepts have a built-in normativity in the sense that their use presupposes something like an epistemic right to expect that whatever objects fall under the rule contain the properties that are expressed in the concept. (p. 442)

Now consider what, following Allison, I will call the 'epistemic luck scenario' (p. 236). On this scenario, the world exhibits a high degree of *de facto* stability: insofar as something is cinnabar it can be counted on to stably instantiate all the properties implied – synthetically and analytically – by that mark. Such a world would seem ordered enough to support and sustain conceptual awareness: the worry of radical disorder which Kant raises in passages such as KrV: A100-1 would be, as a matter of fact, misplaced. Yet the modality of this order is what Allison calls merely 'hypothetical' or 'conditional': if we are to employ concepts, the world must be like *x* and it is indeed like *x*. And this is no different essentially from the following: if we are to breathe, the atmosphere must contain oxygen – and as a happy matter of fact, it does. Now, one of Allison's key claims is that there is something deeply problematic about this story – and that recognising this helps establish the necessity of the categories. This is because the 'epistemic luck scenario':

> calls into question the normativity of empirical cognition by effectively reducing the conformity of appearances to the conditions of apperception to a matter of epistemological luck. (p. 437)

Allison's contention here is effectively that 'empirical concepts/rules presuppose a priori ones' (p. 237): if the lucky scenario is incompatible with empirical cognition, and if empirical cognition is indeed possible, it must be that the categories make it so by guaranteeing some stronger form of order than the de facto stability which the scenario envisages. So when I make an assertion about cinnabar:

> the essential point is that the empirical truth of this and similar claims, i.e., judgments of experience, presuppose that nature as a whole (natura formaliter spectata) is governed by universal laws, which ground the possibility (though not the empirical truth) of such projections and warrant requiring the assent of others. Moreover, in order to preclude the contingency that would undermine normativity as such, and not merely the validity of a particular judgment or candidate for empirical law, these universal laws must have a transcendental ground, which we have seen Kant locates in the categories. (p. 443)

 © Mind Association 2016

Elsewhere he frames it in terms of an 'epistemic right', grounded in our conceptualising practice:

> The essential point is that as rules concepts are themselves normative in the sense that they involve a kind of epistemic right to expect that whatever objects fall under the rule contain the properties that are expressed in the concept-rule and to require that other cognizers, who possess the concept, will agree with one's judgment. (p. 237)

Given this, Allison concludes, there is a 'performative, though not a logical impossibility', in regarding 'the uniformity of nature as merely contingent' (p. 443). In short, if we accept the use of empirical concepts, we must further accept the categories which underpin them; the order on which they rely cannot be down to mere luck.

Such arguments are important both because of their immediate claims and because of their broader implications. For example, the 'oxygen' model sketched fits neatly with Guyer's realist account of the a priori. On this account, the aim is to demonstrate a priori limitations on our cognitive apparatus such that we can only encounter objects insofar as they exhibit property *x*: we could thus know a priori that this property will be exhibited by all objects of our experience, and yet the ground for this, our cognitive restrictions, is evidently compatible with their exhibiting *x* quite independently of us (Guyer 1987, p. 363). But my concern is with the categories, rather than idealism, and here Allison's argument faces an immediate worry, one eloquently articulated by Cassam:

> It is, throughout, an unargued assumption of Kant's that there should, in effect, be some guarantee that appearances will fit into the 'connected whole of human knowledge', but it is surely appropriate to wonder whether a guarantee on this matter ought to be sought. There are two distinct questions to be considered in the present context: what must appearances be like if they are to provide a basis for the unity of consciousness? Secondly, why are appearances such as to provide for the unity of consciousness? The first of these, it may be granted, is a legitimate philosophical question, but it is far from clear that the second is a question with which philosophy is obliged to concern itself. (Cassam 1987, p. 370)

To put it another way, the Humean, broadly construed, can agree that were the world radically disordered, empirical cognition would be impossible – but why does this differ from the link between oxygen and life? Why do we need to invoke the categories to *guarantee* causal order? Allison talks of a 'performative' incoherence in this line of thought, but that seems too strong: there is none of the difficulty in understanding the lucky scenario that we encounter with, say, Moore's paradox.

What of our 'epistemic right'? Allison's idea presumably is that since such order is so integral to the operation of our rational capacities, we can demand it of the world. But he faces three challenges. First is it really so integral? Underlying Allison's view is presumably the point that, lacking the strict universality which comes only with necessity, the luck of the lucky scenario

 © Mind Association 2016

is bound to run out some time. Yet almost every human practice functions despite cases of failure or anomaly, and it is hard to see why a picture on which order, being merely de facto, failed in one in a billion cases should be so radically threatening? Of course, if such order failed consistently and radically, that would be another matter – but the Humean can simply agree with this: it is just another way of making the oxygen point. Second, Allison's argument places enormous weight on the distinctive and inflationary account of concept use Kant offers. Is it really plausible that unreflective agents are committed to such a series of metaphysical implications simply by being able to recognise generic properties? Suppose any breakdown in the merely de facto order will arrive only in the far future: is this fact really sufficient to undermine current practices of discourse and concept application? To put the point from a historical angle, Allison's reading damages the Deduction as a transcendental argument. This is because Hume will not accept its premise: he will argue, as Berkeley does, that there is no need to postulate anything like Kantian empirical concepts in the first place – all that is needed is a tendency to associate groups of particular images (Hume 1978:1.1.7.7-8). Of course, Hume may be wrong about that. But it seems reasonable to prefer approaches on which the categories are shown to be necessary for something that is not itself a philosophical hypothesis, but rather a basic datum of any recognisably human awareness of the world: 'no one', Kant states, will fail to recognise the difference between the ship and the house case (KrV: A190/B235). Third, Allison's approach threatens to rely on idealism in a fashion that is surely too quick. My supposed epistemic right that the world behaves is at root an epistemic demand, and it is unclear why the world should (be honour bound to?) satisfy that. Simply responding 'because it is the world as it is for us' is not good enough when the alleged incoherence of the lucky scenario remains far from obvious.

The second issue to be discussed concerns Allison's account of apperception. As he reads it, Kant's claim in passages such as KrV: B131-2 is effectively that there exists a necessary possibility:

> So considered, Kant holds with Leibniz (and against Descartes and Locke) that the 'I think' need not actually accompany all one's representations, while insisting (against Leibniz) that it must be able to accompany them, at least insofar as these representations are to enter into cognition. In short, though denying that every cognitively significant representation must be apperceived, he maintains that it must be apperceivable. (p. 336)

Yet it is unclear to me why we should be so confident. This is motivated partly by familiar post-Freudian concerns, and partly by the fact that Kant's vision of apperception bundles together many features which seem separable: for example, a kind of direct first person perspective, and the sophisticated representational capacity given by concepts. Consider Moran's case of someone who harbours negative beliefs about her father, and yet is able to access these only from a third person perspective, by drawing inferences from her

 © Mind Association 2016

own behaviour and not by considering the facts about the man and his acts (Moran 2001, p. 85). Such representations would be cognitively significant and yet not apperceivable. One might object that the case is unrealistic, or that all Kant needs is the bare possibility that the subject might regain first personal awareness and deliberative control of these states. (Moran himself of course sees the example as a kind of limit case.) But Allison gives us no direct basis for this guarantee. Even granted such a response, this dialectic suggests that, given the widespread nature of implicit biases, the really important issue is not the necessary possibility of apperception, but how and why it is actually operative in particular cases. More broadly, it was at times unclear to me how much work Allison wants apperception to do. Crudely put, the strategic worry is this: the tighter 'the reciprocity between the unity of consciousness and the consciousness of unity' becomes, the less space there is for the relationship to be genuinely mutually explanatory. Yet, as Allison is acutely aware, this link needs to be extremely tight if one is to avoid characterising apperception in a way that reintroduces Cartesian themes at the heart of the Critical enterprise (pp. 275, 340). I am unclear exactly how he ultimately proposes to walk this tight rope.

The third topic I want to discuss concerns a systemic blindspot in the monograph. The last ten years have seen a huge amount of work on Kant and nonconceptualism. This has obvious and direct implications for the structure of KrV. It is also vital if Kant is to avoid certain basic stumbling blocks. Consider a dog trained to run to the corner of the room, navigating past chairs and tables, and to pick up the smaller of two sticks there: the fact that the dog can perceive spatial relations strikes me as more certain than any argument which might emerge from the *Critique*. Yet if Kant is to avoid problems with such a simple case, he must have some way of allowing animals to perceive and intuit, whilst preserving the view that certain forms of spatio-temporal awareness require the categories. Allison deals with some of these matters in passing (for example, pp. 417-8), although it is hard to see what sense he can really make of the animal case (consider, e.g., p. 419). But his treatment is continually impaired by two features. First, he just assumes that the Deduction and nonconceptualism, understood as the view that beings without concepts might have perception or empirical intuition, are incompatible (p. 418) – and that this rules nonconceptualism out of court. (There are complex issues regarding the definition of 'perception' which I cannot discuss here, but the basic point stands.) I agree that were the two really in tension, the latter would have to go. Yet the relationship is much more complex than Allison seems to realise, and both Allais and I have recently argued, on separate grounds, for their compatibility (Allais 2011; Golob 2016a). Second, the book often reads as if the last decade of Kant scholarship never happened. Not only is there no real treatment of nonconceptualism – probably the topic which has attracted more KrV-focussed research articles than any other in this period – there is no interest at all in any of the broader connections established between the Deduction and

 © Mind Association 2016

contemporary philosophy of mind (for example, with respect to the multiple non-equivalent ways of characterising conceptual content or the implications of relationalist as opposed to representationalist theories of perception). There are, of course, many ways of doing the history of philosophy – and there are few scholars alive who have a better claim to set their own course than Allison. But I raise the point because these omissions do create problems: Allison is in effect deprived of a whole range of sophisticated interlocutors who would have challenged or pushed his views in important ways. For example, Allais's groundbreaking 2009 article on nonconceptualism fails even to make the bibliography, as do all of Hanna's, as do the key works on the conceptualist side such as those by Ginsborg. It is simply not good enough, I am afraid, that the only real engagement with Allais's work in a book that is effectively on the relationship between concepts and intuitions should come in a few lines describing her position on idealism as 'understandable, albeit somewhat too facile' (p. 285).

The final topic I want to cover is the role of time in Allison's argument. There are some issues at the micro-level here which are important. For example, the A deduction states that:

> [e]very intuition contains a manifold in itself, which however would not be represented as such if the mind did not distinguish the time in the succession of impressions on one another; for as contained in one moment no representation can ever be anything other than absolute unity. (KrV: A99)

This opaque passage is important because the status of instantaneous perception is closely linked to the Analogies and the key question of how one draws the distinction between a perception of succession and a succession of perceptions. Suppose, for example, that, having seen an image of a roof and then a door, I next simply step back and see the whole house and thus the two together. This seems sufficient to establish that they do indeed exist simultaneously and thus that my initial experience was merely a subjective succession – yet if this is sufficient, the risk is no category is necessary. Of course, one can try to introduce further sceptical worries (am I sure it is the same roof etc.?) but that seems ad hoc. The literature thus contains many complex analyses of A99, trying effectively to avoid the consequence just sketched without committing Kant to an implausible atomism. Allison's view, if I understand it correctly, is that the 'moment' in A99 is in fact a non-existent limit case:

> [T]here is no contradiction for the simple reason that since, ex hypothesi, an instant or moment is not a part of time *nothing* is intuited in it, including an absolute unity. (p. 210, emphasis in original)

This is not the place for a discussion of Kant's philosophy of perception or time-consciousness, but I was unconvinced by the little Allison said on this, and I am unclear how he would deal with the crude 'stepping back' example I just gave. And this brings me to the macro-level. As I read KrV, Kant's ultimate argument for the categories is closely tied to his account of time. Specifically,

 © Mind Association 2016

my view is that the successive nature of conscious awareness makes the representation of certain relations deeply problematic. Some of these relations are themselves temporal – for example the relation of objective succession discussed in the Second Analogy. Others are mereological – for example, the part-whole relations discussed in the Axioms. Kant's first claim is that even the sceptic, even Hume, assumes the ability to represent such relations; his second claim is that the categories are conditions on doing so (Golob 2016b). From a transcendental perspective, this has obvious advantages: unlike Allison's argument from conceptuality, the premises are ones which Hume can be shown to already accept. Allison's approach from this in multiple ways. For example, when he does argue directly from intuition to the categories, he sees the pre-discursive unity of both space and time, rather than time's successive nature, as doing the work (p. 387). I am hesitant about this, partly since it is unclear what such pre-discursive unity amounts to. (We have a good grip, I think, on the sophisticated unities envisioned by something like Newtonian science and on the simple awareness of space and time possessed by animals, but presumably Allison means neither of these.) But, more importantly, I am perfectly happy to embrace the result that, given the central role of time and the fact that the relevant relations are only treated in detail in the Principles, the Deduction is at best a strategic overview for a set of arguments that start earlier and finish later. Allison, in contrast, is officially committed to resisting any such downgrading of the Deduction (p. 425). One of the many virtues of his magisterial account is that by setting out the resources of the Deduction in such detail, we can now better evaluate its role in the actual arguments of the Critique: be it as 'crown jewel' (p.425) or as propaedeutic.*

* I am very grateful to Prof. Allison for discussion of these issues.

References

Allais, Lucy 2009, 'Kant, Non-Conceptual Content and the Representation of Space', in *Journal of the History of Philosophy*, 47: pp. 383–413.

—— 2011b, 'Transcendental Idealism and the Transcendental Deduction', in D. Schulting and J. Verburgt (eds.) *Kant's Idealism* (London: Springer).

Cassam, Quassim 1987, 'Transcendental Arguments, Transcendental Synthesis and Transcendental Idealism', in *The Philosophical Quaterly* 37.

Golob, Sacha 2016a, 'Why the Transcendental Deduction Is Compatible with Nonconceptualism', in D. Schulting (ed.) *Kant and Nonconceptualism* (London: Palgrave Macmillan).

—— 2016b, 'Kant as Both Conceptualist and Nonconceptualist', in *Kantian Review* [vol. 21].

Guyer, Paul 1987, *Kant and the Claims of Knowledge* (Cambridge University Press).

Longuenesse, Beatrice 1998, *Kant and the Capacity to Judge* (Princeton University Press).

 © Mind Association 2016

Moran, Richard 2001, *Authority and Estrangement* (Princeton University Press).

Abbreviations

References are to *Kants? gesammelte Schriften* (Berlin: de Gruyter, 1900- abbreviated as Ak.). For KrV, however, I employ the standard A/B pagination.

KrV	*Kritik der reinen Vernunft* (Ak., vol. 4)
Log	*Logik* (Ak., vol. 9)
NG	*Versuch, den Begriff der negativen Größen in die Weltweisheit einzuführen* (Ak., vol. 2)
Refl	*Reflexion* (Ak., vol. 14-19)
SvF	*Die falsche Spitzfindzigkeit der vier syllogistischen Figuren erwiesen* Ak., vol. 2)
U	*Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (Ak., vol. 2)

King's College London
sacha.golob@kcl.ac.uk
SACHA GOLOB

doi:10.1093/mind/fzw052
Advance Access publication 7 December 2016

Levels of Argument: A Comparative Study of Plato's Republic *and Aristotle's* Nicomachean Ethics, by Dominic Scott. Oxford: Oxford University Press, 2015.

Dominic Scott's *Levels of Argument* examines Plato's *Republic* and Aristotle's *Nicomachean Ethics* from a comparative methodological perspective. Both authors are 'highly self-conscious about method', notes Scott. His aim is to determine how they conceptualize ethics and practical philosophy as disciplines, and what level of precision (*akribeia*) they consider appropriate when one undertakes theoretical enquiries for practical purposes. Both Plato and Aristotle base their ethics on claims about human nature and the nature of reality, but does practical inquiry require a rigorous and systematic investigation into metaphysics, epistemology, and psychology, or can practical philosophy reach its objective with a less precise grasp of the underlying principles in these areas? 'In other words', asks Scott, 'does practical philosophy have to be as rigorous as those disciplines that investigate the assumptions on which it is based?' (p. 2). These methodological questions have received extensive attention from Aristotle scholars over the past decades. (Treatments include Irwin (1980; 1981; 1988; 2000); Nussbaum (1986); Roche (1988); Whiting (1988); Reeve (1992); Anagnostopoulos (1994); Bostock

 © Mind Association 2016

(2000); Achtenberg (2002); Shields (2007); Zingano (2007); Henry & Nielsen (eds.) (2015).) What is distinctive about Scott's approach is his comparative perspective: the book discerningly elucidates the meta-ethical assumptions underwriting Plato and Aristotle's inquiries into virtue and happiness, and reveals points of affinity that have hitherto been underexplored. Through the sharp lens of Scott's incisive study, the works are brought into alignment and their methodological structures exposed.

The result of Scott's analysis may seem surprising, not least to those who take Aristotle's famous parricide at *EN* I 6 to mark a methodological as well as a metaphysical break with Plato. However, Scott sees similarities.

Though Aristotle eschews both Platonic metaphysics and the ideal of meta-ethical precision, Scott thinks his method in the *Nicomachean Ethics* owes much to the method that Plato actually adopts in the *Republic*. This is not because Plato's ambitions for ethical theory are not higher than Aristotle's – Plato clearly thinks a foundational study of the form of the good and a precise account of the nature of the soul (*Rep.* X 611 b-12a) is required to yield a definitive defence of justice. Such 'precise inquiry', exposing the metaphysical bedrock on which political theories rest, will reveal the soul's 'true nature' by means of 'logical argument' (611c) and will anchor all explanations in an understanding of the good itself. But Plato does not in fact pursue this 'longer route' in the *Republic*, argues Scott – it is gestured at, but not pursued. Instead, Plato settles for a weaker standard of argument, which leads via the 'shorter route' announced at 435c-d to a 'decent' argument that is adequate (*hikanôs*, 435d) for Socrates' audience. The psychological theory and the explication of the virtues produced by the shorter route are imprecise, in that they rest on unproven assumptions and empirical observations. While Plato underlines that a precise answer to Glaucon's challenge requires that we make the full ascent and grasp the 'unhypothetical first principle of everything', the actual method followed in the *Republic* is not this lofty dialectic approach, which would anchor our understanding of individual virtues and rules of conduct in the form of the good, but rather a pedestrian approach that derives a defence of justice from unexamined theoretical suppositions, in the manner prescribed by the method of hypothesis in the *Meno*. Plato's defence of justice is thus tailored to an audience that lacks the enlightened *noêsis* of his future guardians. It is, Scott says, practical philosophy conducted entirely from inside the cave (albeit with a restricted glimpse outside the cave when Socrates presents the sun, line, and cave analogies in books V-VII).

This brings Plato's 'non-ideal' method into line with the method Aristotle recommends in the *Nicomachean Ethics*. Though Aristotle thinks the equivalent of a 'longer road' is possible, based on his own, distinctive brand of metaphysics, Scott holds that Aristotle rejects it in his famous methodological remarks (EN I 3 1094b11-27; I 4 1095a30-b8; 1098a33-b8; II 2 1103b34-1104a10; I 13 1102a23-32; see also VII 1 1145b2-7) as a distraction. Given the practical objective of political science, it is superfluous to search for a precise definition

 © Mind Association 2016

of the soul or to linger on metaphysical principles. Just as carpenters don't need to study the right angle with the precision of a geometrician, someone aiming to make people good need not study metaphysics and psychology with the precision of a metaphysician or a psychologist.

Aristotle, then, takes political science to presuppose the rough and ready arguments of the short road, rather than the foundational inquiries of the longer road, since its aim is to produce good order in citizens and states, and since we can succeed in this objective without becoming theoretical psychologists, theologians, or expert metaphysicians. An approximate grasp will do. The *Republic* and the *Nicomachean Ethics* both address future legislators and their advisors, and both concern virtue and happiness. Having dismissed the search for the 'good itself' as a dead end in I 6, Aristotle identifies his subject matter as the 'human good', of which political science is the 'most architectonic science'. It determines which other sciences should be studied in the state, and at what length. Plato's main concern is whether justice controls happiness – Aristotle seemingly casts a wider net, in that he seeks to work out definitions of the whole catalogue of virtues outlined at *EN* II 7. However, since Aristotle thinks that justice *qua* lawfulness is the whole of virtue, the legislator will need to know what justice in general is. Both works, then, are inquiries into the nature of justice, and both defend the just life as happier than the unjust life. The question is how one best studies these topics given the aims of the inquiry: to produce just citizens and just states.

The book is split neatly down the middle, with one half devoted to each philosopher. By comparing the two works side by side, and rubbing them together like Socrates' fire-sticks, Scott lets their methodologies light up.

In part I, Scott examines Plato's method in the *Republic*, and concludes that despite his attempt to ground his defence of the just life in a theory of forms, the metaphysical theory developed in the middle books (V-VII) plays a minimal role in Socrates' actual response to Glaucon's challenge. Chapter 1 considers the 'shorter route' to a defence of justice in *Republic* II-IV; chapter 2 defends the adequacy of the shorter route, while chapter 3 examines the longer route, which would have covered 'the most important subject' (435c). In its place, Socrates gives us the sun, line, and cave analogies, which by his own admission are merely an approximation. In chapter 4, Scott considers the role of metaphysics in the survey of the four degenerate constitutions and souls, as well as in arguments about pleasure (*Rep.* VIII-IX). In chapter 5, 'The Cave Allegory and the Structure of the *Republic*', Scott hones in on Plato's description of the precise and less precise methods of argument not just as 'routes' (*hodoi*), but as circuits or round-trips (*perihodoi*, 504b2).

This talk of methodological 'circuits' is common to Plato and Aristotle. The long Platonic circuit tracks the education and practice of the future guardians: After reaching knowledge of the form of the good at age fifty, the philosopher kings will understand the nature of the virtues and will be ready to make decisions about particulars. This includes knowing which laws

 © Mind Association 2016

and conventions (*nomima*) to preserve and which to introduce into the ideal state. 'But why exactly is this a circuit?' asks Scott. 'In what way does one come back to the point of departure?' The answer, suggests Scott, 'lies in the fact that the guardians of the ideal state will have been brought up in the correct laws and conventions' (p. 84). They start out with true beliefs about the correctness of their *nomima*, dyed into their souls through their upbringing in childhood and youth. At the age of twenty, they embark upon their mathematical studies, before turning to dialectic at thirty. Only by the age of fifty will they have grasped the nature of the good, which constitutes the summit of their ascent. Now that they understand the nature of the good, they can give a precise explanation of the nature of the forms of the virtues. Their *noêsis* allows them to descend to the much more specific level, 'where they will appreciate the correctness of the *nomima* with which they started, and hence be able to "guard and preserve" them' (484d1-2). It is this final point that constitutes a return, claims Scott, and which 'justifies the description of the whole route as a circuit' (pp. 84-5).

Scott observes that since *both* the long *and* the short routes are described by Plato as circuits, we should expect to find a similar structure (ascent to the universal and descent to the particular) in Socrates' 'imprecise' defence of justice, which reaches its apex in the description of justice as psychological harmony at the end of book IV. In other words, not just the educational journey of the future guardians, but Socrates' actual defence of justice, contains an upwards journey towards a principle, and a downwards journey that defends the beliefs with which the inquiry started and explains why they were correct.

It is not obvious, however, where, if anywhere, Plato completes the 'descent' that defends the account of justice as *oikeiopragia* – each part in the state and soul doing its own work – by showing how justice thus understood leads to the preservation of just behaviour which promotes the common good. Scott frames the issue in light of Sachs' (1963) famous question: Why would a Platonically just person abstain from conventionally unjust acts? Why would we never 'catch him red-handed', traveling down the same road as the unjust man? Scott accuses those who offer answers grounded in the metaphysics of the central books of 'rationally reconstructing' Plato's defence in book IV by 'taking arguments and claims made by Socrates for one purpose and reassembling them for another' (p. 60). Rejecting both 'direct' and 'indirect' versions of such reconstructions, Scott holds that the key elements of Plato's answer, though provisional, are already present in books II-IV. Here we get a 'first draft' of the psychological explanation that is completed in the analysis of deficient constitutions and souls in books VIII-IX. Plato's solution to Sachs' problem starts from the state-soul parallel, says Scott. The appetites of the individual who resembles the just city will have been trained from youth through a mixture of gymnastics and music that makes them harmonious and measured, so that, through his upbringing, he will abstain from *pleonexia*. In both, unnecessary desires are restrained by reason. This

 © Mind Association 2016

anticipates the distinction between necessary and unnecessary desires that Socrates draws in book VIII, and the attribution of lawless unnecessary desires to the tyrant in book IX. This psychological solution is 'in a certain sense self-sufficient: it can stand on its own feet as a plausible way of demonstrating the value of justice' (p. 41) without reference to the metaphysics of the central books. Plato argues, in essence, that a Platonically just soul will be dominated by necessary and lawful desires rather than unnecessary and lawless ones, and that this will ensure that he promotes the common good in the way required of future guardians.

Scott's analysis at this point seems persuasive. However, it is worth asking whether Socrates would consider an explication of the jumbled *nomima* of his interlocutors a requirement of the shorter road. While the future guardians will have been brought up in good habits, the same cannot be assumed for your average Joe (or Coriscus, as it were). Now, Plato takes great care to have Socrates praise the outstanding characters of his brothers, Glaucon and Adeimantus. But even so, it is not a requirement that Socrates be able to show that a person with a Platonically just soul will always or even typically act in conventionally just ways. Plato's account is revisionist, not just in prioritizing an agent-based over an act-based perspective, and so the 'downward' movement of the short circuit will not necessarily return us to the conventional starting point. It is perhaps not so surprising, then, that the outline of the shorter circuit is less clearly drawn and requires more excavation on Scott's part than tracing the trajectory of the long circuit.

We may also balk at the description of the shorter route as 'non-theoretical' or 'non-metaphysical'. To modern ears, the argument in books II-IV seems clearly to rely on metaphysical assumptions, such as the Unitarian Assumption (IV 435a6-9) and the Principle of Contraries (IV 436b). However, Scott underlines that these principles remain suppositions – Socrates argues *from* them, not *to* them. And at no point in books II-IV, argues Scott, does Plato appeal to the existence of forms in his defence of justice. The argument which culminates with Socrates description of justice as 'some kind of (*tis*) health' (444d12) is metaphysically innocent as far as the theory of forms is concerned. The only real appeal to forms that Scott discerns outside of the central books appears in the second pleasure argument in book IX, which Scott describes as 'a brief, temporary, and somewhat disorienting glimpse outside the cave' (p. 96). Plato's reference in III (402b9-c8) to the 'forms' that the guardians must know (of temperance, courage, generosity, and magnificence) is really a reference to immanent characters: these are types or modes intermediate between sensible individuals and forms. Any other reference to forms is extraneous to the main argument of the *Republic*.

Why, exactly, does Socrates fail to make more than a fleeting reference to the theory of forms in his reply to Glaucon's challenge? Scott thinks Socrates' modest approach is born of necessity: 'his interlocutors lack the mathematical and metaphysical background to complete the longer circuit in its entirety'.

 © Mind Association 2016

They may be able to follow what Scott calls a 'middle route', which 'follows the basic pattern of the shorter route', while taking a 'glimpse of the metaphysical perspective associated with the longer route'. This 'mezzanine level argument' 'borrows certain claims that only someone actually taking the longer route in its entirety would be in a position to master' (p. 57). Scott's explanation leaves open the possibility that Socrates could have followed the longer route, had Glaucon and Adeimantus been up to the task. This seems questionable. While it is true that the brothers, whatever their virtues, are not yet enlightened guardians, neither is Socrates. In the *Meno*, the hypothetical method is introduced in the face of Socrates' disavowals of knowledge, and in the *Republic*, Socrates disavows knowledge not just at the end of book I (as Scott duly observes), but also at the critical juncture at the start of the discussion of the 'offspring' of the form of the good at 506c: 'Do you think that it is right to talk about things that one does not know as if one knows them?' In the Glaucus passage at the end of book X, he likewise underlines his ignorance of the soul as it is in its pure state. If Socrates in fact conceals knowledge he possesses for the sake of his audience, then Thrasymachus was right in book I to charge him with 'his usual irony' (337a). But there is no reason to suppose that Socrates is being coy when he declares that he lacks knowledge of the good. On this interpretation, the *Republic* is a less dogmatic work than critics often allow – like the 'transitional' dialogues identified by developmentalists, it displays an agnostic Socrates who waves his hand at the forms, stating roughly and in outline what they are *like*, without giving a definition of their essence.[1]

Having defended his novel and provocative reading of the *Republic*, Scott in the second half of the book turns to Aristotle, and his famous methodological remarks about precision. Having helpfully distinguished between what he calls 'practical' and 'theoretical' precision – the first of which concerns normative theory's inability to articulate context-independent precepts that a political scientist could apply mechanically – Scott concentrates on theoretical imprecision. This does not concern the application of normative principles, but rather the meta-ethical and metaphysical presuppositions of the principles that Aristotle uncovers in the *Nicomachean Ethics* – specifically, the principles he alludes to in his project of spelling out the content of the happy life. This project takes up the whole of the *Ethics*, and requires an elucidation of the virtues, of the grounds of praiseworthiness and blameworthiness, and the nature of pleasure and friendship, among other topics. As Scott reads him, Aristotle denies that political science requires an extensive understanding of the metaphysical, psychological, epistemological, or theological principles undergirding the definitions he outlines. Though it is possible to give a precise explication of the principles behind e.g. the function argument and

[1] As Lesley Brown reminds me, it may be that the fictional Socrates avoids the longer route because his author, Plato, has not yet worked out the details of the theory of forms.

 © Mind Association 2016

the division of the soul, would-be experts in political science need not master them to succeed. Consequently, those who seek support for ethical claims in the *Metaphysics*, *Physics*, or *De Anima*, ignore Aristotle's warnings about precision – they fail to respect Aristotle's conception of ethics as an imprecise science. This is not to deny that a distinctively Aristotelian 'longer road' is possible, Scott argues, but it is a distraction for students of political science: 'In complete contrast to Socrates in *Republic* 504b-e, [Aristotle] does not think that achieving such precision would do anything to enhance one's ability to make correct decisions in the practical arena' (p. 140).

Scott here rejects positions defended by Irwin, Bostock, Whiting, and Shields, who maintain that Aristotle bases the theory of the *Nicomachean Ethics* on those developed in his other works, and that mastering these works is necessary for the political scientist. Scott's own position is open to interpretation, however. It is one thing to say, as Scott initially does, that the exact study of psychology does not 'belong to' political science, and quite another to claim that political science can get along just fine without a precise study of e.g. psychology. As Scott later allows, psychology could be foundational for ethics without being part of it as a discipline, just as cell biology is distinct from oncology, but still foundational for it.

The question, then, is whether Aristotle believes that the 'longer route', via the more foundational disciplines, is an appropriate route to travel for the intended readers of the *Nicomachean Ethics*. Here, Scott maintains that we should distinguish sharply between rational reconstructions and explications of Aristotle's own recommendations. Whom does Aristotle warn against excessive theoretical precision and why? Pace Scott, one might object that Aristotle needn't be cautioning philosophers against studying psychology or metaphysics as part of their overall curriculum, but only against incorporating the precise study of psychology or metaphysics into the study of ethics, as Plato attempts in the *Republic*. That 'integrated' approach would indeed result in digressions 'crowding out' the main issues (*ta parerga tôn ergôn pleiô ginêtai*, *EN* I 7 1098a33), and so dividing up the inquiries into separate lecture courses is the way to go. By contrast, Scott takes Aristotle to argue that ethics and politics can be studied successfully without more than a quick glance at the underlying metaphysical and psychological theories. This is not simply because its objective is practical rather than theoretical – Aristotle thinks that a theoretical treatise into the human good will aid the lawgiver – but because the extra-ethical theoretical requirements for political prudence are limited. While arguing that Aristotle's audience consists of 'philosophers familiar with these more exact areas of inquiry' (p. 140), Scott envisions Aristotle cautioning them that they must 'put their usual desire for precision on hold and orient themselves in a more practical direction' (p. 140) when they study political science as future advisors to legislators. But if they have already studied the *De Anima*, *Posterior Analytics*, and *Physics* before turning to practical philosophy, as some of Aristotle's allusions would suggest, it

 © Mind Association 2016

seems quaint to say that they put their usual desire for theoretical precision 'on hold'. Scott's interpretation would have made better sense had he thought the *EN* was aimed at the legislators themselves rather than their tutors – though even here, Aristotle says that politicians must 'study the soul' just as a better sort of eye doctor must study the whole body and not just the eyes alone (*EN* 1102a17-19; 23-4; cf. *Pol.* 1333a37; see Shields (2015)).

In attributing to Aristotle the view that it is *possible* to find further explanations for the principles of political science, but *superfluous* given its practical nature, Scott allows that those who read the *Nicomachean Ethics* may learn something valuable by reconstructing the theoretical underpinnings of Aristotle's ethics. Few scholars nowadays read the *Nicomachean Ethics* as a 'mirror for princes' – whether to produce virtuous legislators or to avoid tutoring a future George Osborne or Boris Johnson. Aristotle may even think that we *should* aim for precision, just as students in the Lyceum were expected to complete a full course of study, comprising all branches of inquiry. What they cannot do is study foundational sciences in detail as part of a lecture course on ethics. In other words, the Oxbridge-approach, where the *Republic* and *Nicomachean Ethics* are studied by students with no grounding in their respective metaphysics or epistemology, may be at best superficial, and at worst pernicious, a kind of dabbling that can't do justice to the strength of Aristotle's position, while leaving students ill equipped to defend Aristotle's views against e.g. relativist or anti-naturalist critics. If Aristotle accepts the reading that Scott proposes, he arguably underestimates the normative implications of accepting the right metaethical theory – realists and conventionalists tend to adopt different normative principles. Unlike Scott, I take Aristotle to hold that the 'long route' would also benefit Aristotle's ideal legislator: though the purpose of ethics is to become good, we're more likely to succeed if we know what the human good is, and since the good for any living organism is relative to its nature, the study of ethics presupposes the study of psychology. His warnings about precision tell us to put aside questions about foundations, since these cannot be treated adequately in a lecture course on ethics. And in any case, they are covered at length elsewhere.

In chapter 9, Scott considers a methodological passage in *EN* I 4, where Aristotle describes his own inquiry with reference to a puzzle Plato 'used to raise':

> We must notice, however, the difference between arguments from principles (*archai*) and arguments towards principles. For indeed Plato was right to be puzzled about this, when he used to ask if [the argument] set out from the principles or led toward them – just as on a race course the path may go from the starting point to the far end, or back again. For we should certainly begin from things known, but things are known in two ways; for some are known to us, some known without qualification. Presumably, then, we ought to begin from things known to us. (*EN* I 4 1094b32-1095a6)

To appreciate Aristotle's race-course analogy, recall that the track of a Greek stadium leads in a reclining U-shape from the starting point to the far end,

 © Mind Association 2016

and back again. Aristotle compares the first leg leading up to the mid-point to the movement *towards* first principles, and the leg leading back down to the movement *from* first principles.

Aristotle maintains that we 'need to have been brought up in good habits' if we are to be adequate students of fine and just things, and of political questions more generally. 'For', he says, 'we begin from the [belief] that [something is true]; if this is apparent enough to us, we can begin without also knowing why [it is true]'. Only someone who is well brought up has the beginnings, or can easily acquire them. Someone who hasn't been brought up in good habits, and who lacks the ability to acquire the starting points by himself, should listen to the advice of others: for he who 'neither grasps it himself nor takes to heart what he hears from another is a useless man' (EN I 4, 1095b10). The aim of the *Nicomachean Ethics* is to articulate the views from which we must start, and to use them as step-stones in the definition of the human good. The movement from 'the that' (*to hoti*, the beliefs with which we start) to the first principle (*archê*), namely the account of *eudaimonia*, allows us to return to our starting points and to explain why our beliefs are (mostly) right, and, in cases where a mistake has been revealed, why people were prone to error. The first principle thus provides 'the because' (*to dihoti*), or justification, of things better known to us – the assumptions from which the inquiry starts. If Scott is right, Aristotle's nod to Plato is not a coincidence. This 'ascent and descent'-structure also marks the trajectory of Socrates and his interlocutors in *Republic* II-X, as Socrates suggests at *Rep.* 510b-c and 533c-d, when he outlines the dialectical mode of inquiry.

Scott now asks: If Aristotle shares Plato's commitment to the upward and downward journey of inquiry, where in the *Nicomachean Ethics* does he make the descent to show that the first principles uncovered in his definition (*horismos*) of happiness and virtue can explain the opinions from which the journey towards first principles started? Though Aristotle 'envisages' his inquiry taking both journeys, in fact, argues Scott, 'almost all' of the *Nicomachean Ethics* is exhausted by one of the journeys described in the race-course analogy (p. 173), namely the specification of the components of *eudaimonia*. He never gets around to explaining the intuitions. Had Aristotle intended to make the reverse journey in the *Nicomachean Ethics*, argues Scott, 'we should expect to find a clearly marked point in the work where he has completed his account of the principle and then starts to move in the reverse direction. I can find no such point, and no signposts to suggest such an interpretation' (p. 175). Scott considers, but ultimately rejects the proposal that the *Politics* might be the sequel that completes the downward journey. I am at a loss to understand why Scott assumes there must be *one* marked point, however. Scott overlooks another way in which Aristotle may complete both the upward and downward journeys, namely many times, once for each topic that he introduces in the *Ethics*. In a recent paper, Daniel Harkin (2015) has discerned a three-phase pattern in Aristotle's discussion of individual

 © Mind Association 2016

virtues which explains each individual analysis as a lap around the Aristotelian race course. Aristotle starts by giving an account of the domain of the virtue (its '*peri ho*'); in the next phase, he identifies the common denominators in the *endoxa*, and gives a mixture of explanatory and merely generalizing '*hôs epi to polu*' statements, which yield a determinate account of the virtue – the definition. Reversing the dialectic, in the final phase he then proceeds from the definition to a corrective discussion of the *endoxa*. This pattern is exemplified in Aristotle's discussion of courage, generosity and temperance; even the otherwise puzzling discussion of magnanimity follows the same route. Had Scott considered the possibility of a circuit internal to the discussion of each virtue, he would not have come up empty-handed. The pattern recurs in Aristotle's discussion of happiness in book I, and also, arguably, in his discussions of friendship and pleasure.

That Scott never (explicitly) considers this possibility is puzzling, but perhaps understandable in light of his resistance to the idea, defended by Barnes (1981) and others, that the so-called method of *endoxa* outlined in *EN* VII 1 is 'the' method of the *Nicomachean Ethics*. In the final chapter of *Levels of Argument*, Scott maintains that the method of *endoxa* should not be assimilated to the one described in the racecourse analogy: 'The reality is that there is no one method of ethics to be found in the *EN*' (p. 209). However, there is no need to assimilate the inquiry to and from principles in individual chapters to the endoxic method as described in VII 1; whether the analyses of happiness, the virtues, friendship, and pleasure follows the recipe specified in this common book is a separate issue from whether he follows the trajectory of the race course. The literature on the endoxic method is too extensive to allow a responsible assessment of Scott's reading in this (relatively) brief review, but again, Scott's analysis helps bring important questions about Aristotle's method into sharp focus; he assesses the evidence judiciously, and defends plausible positions in a way that allows the reader to make up her own mind. In terms of style, Dominic Scott is perhaps the most elegant writer working on ancient philosophy today. *Levels of Argument* sets a new benchmark for scholarship. It is one of the most original and thought-provoking works of ancient philosophy to appear in recent years. Like the works it studies, it rewards reading and re-reading, and will become a standard critical work for any serious student of ancient philosophy.

References

Achtenberg, D. 2002, *Cognition of Value in Aristotle's Ethics* (Albany: SUNY Press).

Anagnostopoulos, G. 1994, *Aristotle on the Goals and Exactness of Ethics* (Berkeley: University of California Press).

Barnes, J. 1981, 'Aristotle and the Methods of Ethics', *Revue Internationale de la Philosophie* 34, pp. 490–511.

Bostock, D. 2000, *Aristotle's Ethics* (Oxford: Oxford University Press).

 © Mind Association 2016

Harkin, D. 2015, 'How to Solve the Problem of *Megalopsuchia*' (unpublished MS).
Henry, D. and Nielsen K. M. (eds.) 2015, *Bridging the Gap between Aristotle's Science and Ethics* (Cambridge: Cambridge University Press).
Irwin, T. H. 1980, 'The Metaphysical and Psychological Basis of Aristotle's Ethics', in A. Rorty (ed.), *Essays on Aristotle's Ethics* (Berkeley: University of California Press), pp. 35—53.
——1981, 'Aristotle's Methods of Ethics', in D. J. O'Meara (ed.), *Studies in Aristotle* (Washington DC), pp. 193–224.
——1988, *Aristotle's First Principles* (Oxford: Oxford University Press).
——2000, 'Ethics as an Inexact Science' in B. Hooker and M. Little (eds.), *Moral Particularism* (Oxford: Oxford University Press), pp. 100–29.
Nussbaum, M. 1986, *The Fragility of Goodness* (Cambridge: Cambridge University Press).
Roche, T. 1988, 'On the Alleged Metaphysical Foundation of Aristotle's Ethics', *Ancient Philosophy* 8, pp. 49–62.
Reeve, C. D. C. 1992, *Practices of Reason: Aristotle's Nicomachean Ethics* (Oxford: Oxford University Press).
Sachs, D. 1963, 'A Fallacy in Plato's Republic', *Philosophical Review* 72, pp. 141–58.
Shields, C. 2007, *Aristotle* (Abingdon and New York: Routledge).
——2015, 'The Science of Soul in Aristotle's Ethics', in Henry and Nielsen (eds.), pp. 232–53.
Whiting, J. 1988, 'Aristotle's Function Argument: A Defense', *Ancient Philosophy* 8, pp. 33–48.
Zingano, M. 2007, 'Aristotle and the Problems of Method in Ethics', *Oxford Studies in Ancient Philosophy* 32, pp. 297–330.

Somerville College & Faculty of Philosophy, KAREN MARGRETHE NIELSEN
Oxford University
Woodstock Road
Oxford OX2 6HD
karen.nielsen@some.ox.ac.uk
doi:10.1093/mind/fzw054 Advance Access publication 23 December 2016

Proof-Theoretic Semantics, by Nissim Francez. London: College Publications, 2015. Pp. xx + 415.

1. Introduction

After years of study of formal semantics and computational linguistics, Nissim Francez took up the gauntlet thrown down by Schroeder-Heister (Kahle and Schroeder-Heister 2006) regarding the expansion of proof-theoretic semantics

 © Mind Association 2016

(PTS) towards natural language. A series of papers devoted to this issue culminated last year in a book, which was published by College Publications. As the author states in the preface, 'the book attempts to present as a coherent story both the main line of research pertaining to the PTS in logic, as well as the extension of PTS to NL [natural language] – an ongoing endeavour' and provide a 'reference resource on PTS' (Francez 2015, p. vii).

First, I will briefly go over the contents of the book. Then I will offer my general assessment.

2. Contents

The book is split into two main thematic blocks dedicated to PTS in logic (roughly three fifths of the book) and in natural language (two fifths of the book), respectively. These two main parts are accompanied by an opening 'Introduction' section (one fifth of the book), which familiarizes the reader with the general context of the work as well as with the terminology and tools necessary for the rest of the book. Only basic acquaintance with model-theoretic semantics is assumed.

The first part is called 'Proof-Theoretic Semantics in Logic' and it surveys mainly the works of others, but the author's new contributions are presented as well, most notably, the ideas of explicit definition of 'reified' PTS meanings and canonical derivations from open assumptions.

The book covers all the standard PTS topics that the potential reader might be curious about (and even some of the less common ones), for example, different views of PTS (I-view, E-view ...), questions surrounding the PTS notions of harmony (validity, canonicity, assumptions ...), matters concerning appropriate systems (intuitionistic vs. classical), type of rules (natural deduction vs. sequent calculus, and single conclusion vs. multiple conclusion), the nature of negation (unilateralism vs. bilateralism), etc.

The second part is called 'Proof-Theoretic Semantics for Natural Language' and, as the slight change in title suggests, this is where the majority of novel research is located. As hinted at earlier, PTS approaches to natural language are still in their infancy (despite large strides being taken recently by, for example, Francez, Więckowski 2011, 2016 and Chatzikyriakidis and Luo 2014). The author hopes that this part will establish the 'viability of this approach [PTS] as an alternative to the model-theoretic semantics, in which, so it seems, most linguistics semanticists are entrenched' (Francez 2015, p. vii).

To name the most prominent topics covered, Francez offers a PTS account of the meaning of sentences with transitive/intransitive verbs and sub-sentential phrases, determiners, intensionality and contextual meaning variation.

3. Opinion

Plentiful references, illustrative examples and overall textbook-like exposition make this book a great resource for both PTS beginners as well as veterans.

 © Mind Association 2016

From this perspective, Francez successfully fulfils the task he set for himself. A book such as this was indeed long time coming – it provides a good gateway for any reader interested in PTS. Perhaps more space could have been dedicated to a discussion regarding the relation between PTS and model-theoretic semantics, but I understand the author's decision to keep it brief, as it is probably better left for a more focused publication.

Restall in his recent review (2016) of *Advances in Proof-Theoretic Semantics* (2016), the first volume in the Springer *Trends in Logic* series devoted solely to PTS, wrote:

> The authors of papers in this volume (almost all) agree that proof theory is important to semantics, and they mostly agree that this has something to do with natural deduction and the sequent calculus, but beyond that, there is little agreement on the important problems to tackle or on which framework might best be used to tackle those problems. PTS in this narrower sense is very much a field in flux, and this volume shows it.
>
> [...]
>
> There are many ideas and approaches on display, with some shared themes and techniques, but there is not so much in the way of coherent and developed methodology. I can only hope there is more to come in this tradition, because there are connections to be made between meaning, use, rules and inference. (Restall 2016)

While I agree with many of Restall's points, I think a lot of the things he is looking for could be found in Francez's book. More specifically, with its systematic approach and carefully selected set of tools it offers a uniform framework that can accommodate many of the existing PTS approaches (unilateral, bilateral, I-canonicity, E-canonicity ...) and provide much more (for example, a treatment of natural language). This goes together with the author's own approach: '... [I] rather view PTS *in a variety of forms* as a universal *definitional tool*' (Francez 2015, p. viii, emphasis in original).

But I do not want to reduce this book to a good PTS handbook. As mentioned before, while the first half of the book is dedicated mainly to a systematic study of already existing material, it is the second part that contains most of the original research – based almost entirely on the previous work done by the author and his collaborators, namely Dyckhoff and Ben-Avi.

The central notion of 'reified' PTS meaning, upon which Francez's system stands not only allows for an elegant technical treatment of other PTS notions, but also has practical consequences (e.g., compatibility with type logical grammar), although I imagine that not all inferentialists will be thrilled by the idea of seeing meaning once again as some abstract entity (even though proof-theoretic in its nature). Nevertheless, it is hard to envision how the enterprise of formal semantics can be reasonably carried out without such hypostatized theoretical constructs.

I especially appreciated the parts about the granularity of PTS meaning, demonstrating that Francez's system is truly hyperintensional (according to

 © Mind Association 2016

Cresswell's 1985 original specification), since it can distinguish between logically equivalent propositions. At the same time, Francez also proposes simple remedies about how to restrain this unbounded hyperintensionality, utilizing the notion of grounds of an assertion.

On the other hand, a PTS treatment of the meaning of atomic propositions is missing. As is emphasized on many occasions (see e.g., Francez 2015, pp. 196, 245, 258, 259, 314), the meaning of atomic propositions is assumed to be given externally to the meaning-conferring PTS system. However, the author is well aware of this 'blind spot' and at the very end he frames it as the next open challenge for PTS. (More on this in §4.)

Another topic that should be brought up is that philosophical issues remain in the background throughout the whole book. And this is by design and telegraphed by the author at the very beginning:

> There is an attempt in this book to downplay the role of philosophical arguments and disputes, such as realism vs. anti-realism, underlying many of the discussions of PTS, and concentrate instead on the technical details involved in developing PTS. (Francez 2015, p. vii)

Thus, to readers interested in this subject matter, I would recommend that they look elsewhere, e.g., Dummett's *The Logical Basis of Metaphysics* (1991), Brandom's *Making It Explicit* (1994) or *Articulating Reasons* (2000), or Peregrin's *Inferentialism* (2014). Although these books do not deal with PTS (in Dummett's and Brandom's books it is not mentioned at all and in Peregrin's book it appears only as a short side note), they cover a lot of topics related to its background philosophy and fundamental principles.

Of course, it is an open question how necessary these issues are for the advancement of PTS and its successful deployment in practice. That is not to say, however, that all philosophical topics are unimportant; there are still a lot of avenues for further investigations. (See e.g. the list of open problems for PTS, both technical and philosophical, at the end of this review.)

4. Conclusion

The book shows not only that PTS approaches to natural language are feasible, but that they carry advantages as well. For example, I think there is a lot of untapped potential in regards to its built-it hyperintensionality, i.e., fine-granularity of meaning.

Consider e.g. the following Stalnaker quote:

> …[W]e lack a satisfactory understanding, from any point of view, of what it is to believe that A while disbelieving that B, where the A and the B stand for necessarily equivalent expressions. (Stalnaker 1984, p. 24)

On the PTS account, there seems to be no innate mystery: to believe A and not B, where A and B are logically equivalent, is simply to be 'convinced' by the proof of A but not by the proof of B (or more precisely, to be able to recognize the proof of A but not the proof of B). There appears to be nothing

 © Mind Association 2016

puzzling about this position given the varying complexities of proofs or, more generally, justifications.

And speaking of Francez's system specifically, Luo (2014) recently also pointed out the possibility of embedding it in his own PTS-like system (based on Martin-Löf's 1984 constructive type theory). This is certainly another interesting route, the perusal of which has the potential to give rise to new and exciting possibilities.

Before closing I would like to point out several directions for further research that Francez proposes throughout his book:

1. PTS treatment of modal operators (Francez 2015, p. 115)
2. Relating multiple-conclusion natural-deduction to actual inferential practices (Francez 2015, p. 149)
3. Extending PTS towards multi-sentential linguistic constructs (dialogues, discourses) (Francez 2015, p. 322)
4. PTS meaning of atomic sentences (leading to Proof-Theoretic Lexical Semantics) (Francez 2015, p. 377)

Francez explicitly adds the last one to the three challenges for PTS formulated by Schroeder-Heister (Piecha and Schroeder-Heister 2016, p. 253):

5. The nature of hypotheses and the problem of the appropriate format of proofs
6. The problem of a satisfactory notion of proof-theoretic harmony
7. The problem of extending methods of PTS beyond logic

And if we add also those 'unofficial' ones mentioned by Restall (2016), we get:

8. The challenge of better exploring interface and connections between natural language (and real-life deductions) and abstract proof structures
9. The challenge of extending PTS beyond the semantics of assertions towards different kinds of speech acts
10. The challenge of investigating more connections between meaning, use, rules and inference
11. The challenge of examining in more depth how the notions of proof are connected to language use and understanding

To conclude, PTS still has a long road ahead, as indicated by the above list. But without any doubt it is also a promising area of novel research, as clearly demonstrated by Francez's book.

References

Brandom, Robert 1994, *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge, MA: Harvard University Press).

 © Mind Association 2016

——2000, *Articulating Reasons: An Introduction to Inferentialism* (Cambridge, MA: Harvard University Press).
Chatzikyriakidis, Stergios and Zhaohui Luo 2014, 'Natural Language Inference in Coq', in *Journal of Logic, Language and Information* 23(4).
Cresswell, Max John 1985, *Structured Meanings* (Cambridge, MA: MIT Press).
Dummett, , Michael 1991, *The Logical Basis of Metaphysics* (Cambridge, MA: Harvard University Press).
Kahle, Reinhard and Peter Schroeder-Heister 2006, 'Introduction: Proof-Theoretic Semantics', in *Synthese* 148(3).
Luo, Zhaohui 2014, 'Formal Semantics in Modern Type Theories: Is It Model-theoretic, Proof-theoretic, or Both?', in N. Asher and S. Soloviev (eds.), *Logical Aspects of Computational Linguistics: 8th International Conference, LACL 2014, Toulouse, France, June 18–20, 2014. Proceedings* (Berlin, Heidelberg: Springer Berlin Heidelberg).
Martin-Löf, Per 1984, *Intuitionistic Type Theory* (Napoli: Bibliopolis).
Peregrin, Jaroslav 2014, *Inferentialism: Why Rules Matter* (London: Palgrave Macmillan).
Piecha, Thomas and Peter Schroeder-Heister (eds.) 2016, *Advances in Proof-Theoretic Semantics* (Cham: Springer).
Restall, Greg 2016, 'Advances in Proof-Theoretic Semantics', in *Notre Dame Philosophical Reviews* <http://ndpr.nd.edu/news/67011-advances-in-proof-theoretic-semantics/>.
Stalnaker, Robert 1984, *Inquiry* (Cambridge, MA: MIT Press).
Więckowski, Bartosz 2011, 'Rules for Subatomic Derivation', in *Review of Symbolic Logic* 4(2).
——2016, 'Subatomic Natural Deduction for a Naturalistic First-Order Language with Non-Primitive Identity', in *Journal of Logic, Language and Information* 25(2).

Department of Philosophy,
Faculty of Arts,
Masaryk University
ivo.pezlar@gmail.com
IVO PEZLAR

doi:10.1093/mind/fzw056 Advance Access publication 17 November 2016

Strange Tools: Art and Human Nature, by Alva Nöe. New York: Hill and Wang, 2015. Pp. xiii + 285.

Like all of his work, Alva Nöe's latest book *Strange Tools* is written with passion and intelligence. It is fresh and incisive, and defends the fascinating idea that art and philosophy play similar rôles in our lives. According to Nöe, they are both essentially reflective or critical activities. Each of the arts reflects

 © Mind Association 2017

on, 'displays', and 'investigates', a distinctive human activity, such as story-telling, dancing, or picture-making, in roughly the way that philosophy, as Nöe sees it, reflects on science.

This is a doubly unorthodox position. It combines an unorthodox view about the arts with an unorthodox view about philosophy.

First, regarding the arts, the orthodox view is that epic poetry *is* a kind of story-telling, ballet *is* a kind of dance, and painting *is* a kind of picture-making. The artistic activity may be especially preoccupied with, and therefore especially sensitive to, the aesthetic potential of its medium, perhaps even to the point where its original practical function is neglected or set aside. But although the *Iliad* is a different *kind* of narrative from a military dispatch, it is still a narrative; and although Balanchine's *Serenade* is a different kind of dance from the All Blacks' Haka, it is still a dance.

Nöe rejects this traditional conception of the arts. He distinguishes between two levels of thought and activity:

> Level 1 is the level of the organized activity or the technology. Level 2 is the level where the nature of the organization at the lower level gets put on display and investigated. At level 1, we have activities like talking, moving, dancing, making pictures, singing, etc. ... Correspondingly, at level 2, we have the different arts: poetry and fiction, choreography, painting and photography, music, and so on. Level-2 practices play with and reshape level-1 activities. (p. 29*f*)

For example,

> When a choreographer stages a dance, he is representing dancing. That is, he puts dancing itself on display. Choreography shows us dancing, and so, really, it displays us, we human beings, as dancers; choreography shows us dancing; choreography exhibits the place dancing has, or can have, in our lives. (p. 13)
>
> Choreography [...] is not dancing, it is an engagement with dancing as a phenomenon. (p. 15)

Turning to philosophy, according to one traditional view, philosophy is the most fundamental, or alternatively the most general, part of science. For example, in Descartes' famous image, which appears in the preface to the *Principles of Philosophy*, the whole of science is like a tree: 'The roots are metaphysics, the trunk is physics, and the branches emerging from the trunk are all the other sciences'. In the twentieth century, Russell held that the propositions of logic are supremely general truths about the most pervasive features of reality, while Sellars made the vague but much-quoted claim that the aim of philosophy is 'to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term'.

Again, Nöe rejects this conception of philosophy. He prefers the view that philosophy is a critical or reflective activity. This is not as radical a departure from tradition as Nöe's conception of the arts. In modern philosophy, Kant is its pre-eminent exponent, although Nöe's version of it owes more to Wittgenstein.

 © Mind Association 2017

'Philosophy', Nöe insists, 'is not just "more science"' (p. 220). '[It] doesn't yield positive nuggets of information that you can take away and put to work in this or that area of your life, the way physics, mathematics, or economics does' (p. 115). He quotes Wittgenstein's remark in the *Tractatus*: 'The word "philosophy" must mean something which stands above or below, but not beside the natural sciences' (4.111).

Putting these ideas about the arts and philosophy together, Nöe's view is that they are similar kinds of activity:

> Philosophy stands to ... reasoning, argument, belief formation, and, crucially, the work of science in the same kind of relation that, say, choreography stands to movement and dancing, or painting as an art stands to picture-making activities as these flourish in our lives. (p. 29*f*)
>
> Both philosophy and choreography aim at ... a kind of understanding that, in Wittgenstein's phrase, consists in having a perspicuous representation —but they do it, so to speak, in different neighborhoods of our existence. (p. 17)

I have some sympathy both with Nöe's conception of the arts, and with his conception of philosophy. I shall comment on them in turn.

First, regarding art, the distinction between level 1 activities and level 2 activities does not seem to me helpful, and to be fair, Nöe himself expresses some unease about it. He says that 'the idea that choreography is metadancing, or that art practices are, as I have been suggesting, metalevel, is too simple' (p. 30). Whatever else figurative painters do, they certainly make pictures; and whatever else Homer did, he did tell stories. So the idea that their work reflects on picture-making without being picture-making, or reflects on narrative without being narrative, cannot be right.

But Nöe *is* right to suggest that artists, at least the artists who shape the artistic traditions to which they belong, are bound to reflect critically on the medium in which they work, and often challenge established ideas about its potentialities and limitations. And this is not an accidental feature of the arts.

Every work of art communicates thoughts, feelings, and perceptions in a specific medium, with specific materials and techniques. And its interest and value as a work of art always depend on its sensitivity to the communicative possibilities of the medium—of the materials and techniques that it employs. So the practice of making art, at least when it is imaginative to some degree, necessarily involves reflecting on the nature of the medium, and on its expressive potential. This isn't just a feature of the subversive art of the twentieth century, it is a feature of art in every tradition, because it is part of what makes a practice of making pictures, or narratives, or dances, count as art.

(This reflective or reflexive nature is one important difference between art and science. For example, physics in the seventeenth century *was* simultaneously engaged in searching for the laws of motion and in searching for the right way to search for the laws of motion. But this isn't an essential and universal part of physics: physics since the eighteenth century hasn't had this dual nature.)

 © Mind Association 2017

Nöe is so concerned to underline the importance of this aspect of art that he sometimes suggests that the whole purpose and meaning of art consists in this reflective and critical attitude to the artistic medium. 'A work of art', he says, using a phrase that gives the book its title, 'is a strange tool; it is an implement or instrument that has been denuded of its function. Art is the enemy of function, it is the perversion of technology. This is why architecture has a problematic standing among the arts' (p. 98).

This is surely an exaggeration. First, every, or almost every, work of art has a communicative function, and many works of art have other functions as well. So art is not an enemy of function, in any plausible sense of the phrase. Second, it is true that there is nothing intrinsic about buildings that makes them works of art. (If the Vikings had done aesthetics, architecture would not have been included in their conception of the arts, but shipbuilding would.) But there have been times in the history of every civilization when buildings have been the most important artefacts that express values in a sensuous and symbolic form, and when painting and sculpture have been subordinated to architecture. So a theory of art that implies that architecture has a problematic standing among the arts is a problematic theory of art.

However, despite the exaggeration, works of art *are* strange tools, because they combine a communicative function with a peculiar absorption in the medium in which they communicate. The most poignant image of this absorption, of its profound and disturbing power, is Shakespeare's image of the dyer's hand:

> [...] my nature is subdued
> To what it works in, like the dyer's hand:
> Pity me, then, and wish I were renewed.

Turning to philosophy, again I agree partly with Nöe's view. That is, I agree that philosophy is not the most fundamental or the most general part of science. And I agree that philosophy is a reflective and critical discipline. But unlike Nöe, I do not want to deny that philosophy is an integral part of natural science, just as I do not want to deny that painting *is* picture-making, and epic poetry *is* narrative.

Nöe contrasts philosophy on the one hand with physics and mathematics on the other. As mentioned above, he says: 'Philosophy is not just "more science"'. '[It] doesn't yield positive nuggets of information that you can take away and put to work in this or that area of your life, the way physics, mathematics, or economics does'. I think this is a mistake. In fact one good way to understand the place of philosophy in science is to compare philosophy and mathematics.

Is mathematics part of science? It is not hard to understand the reasons for saying no. Mathematical theorems are not proved or disproved in the way that scientific theories are confirmed or disconfirmed, by conducting experiments and making observations—with telescopes, microscopes, spectroscopes, or the

 © Mind Association 2017

naked eye. Besides, mathematics is as deeply involved in other aspects of our lives as it is in science—in finance and gambling, for example.

Nevertheless, there is an obvious sense in which mathematics is an indispensable part of science. Modern science is unimaginable without mathematics. Physical theories are formulated in mathematical terms, and the observations that confirm or disconfirm them are expressed in numbers. And mathematics is also part of science in a more general, cultural sense, because it is pursued for the sake of enlarging knowledge and understanding, independently of the ways in which it can be used to solve practical problems.

However, the distinctive rôle mathematics plays in science has not always been properly understood. For example, it is a mistake to conceive of geometry, as Descartes did, as the most general science of matter, or to believe, as Kant did, that geometry and arithmetic are a priori bodies of knowledge about space and time. Mathematics is not one of the special sciences; it is not the most fundamental science; and it is not the most general science. Its special task is to explore, extend, codify, and correct all of our thinking and reasoning about quantities and magnitudes. So it permeates every part of science where this kind of thinking and reasoning occur, which today means every part of science. *The Origin of Species* and *The Descent of Man* may be the last scientific masterpieces which do not include any mathematical reasoning or mathematical ideas.

In my view, the relationship between science and philosophy is similar to the relationship between science and mathematics. Mathematics is and always has been largely constructive, whereas philosophy is and always has been largely analytical and critical. In other words, philosophy aims on the whole to understand, criticize, and reform existing systems of concepts rather than to devise new ones. But like mathematics, philosophy is an a priori discipline. And like mathematics, philosophy has made an indispensable contribution to the development and revision of scientific ideas. For example, Special Relativity is just as dependent on Einstein's revolutionary analysis of the concept of simultaneity as it is on nineteenth century developments in geometry. Self-conscious philosophy, deliberately focused on the analysis and criticism of concepts, probably began with ethical and religious concepts. But wherever it begins, it is bound to ramify through our intellectual lives, and scientific concepts—together with ethical, logical, and psychological concepts—have been at the heart of philosophy throughout the modern period.

So, returning to Nöe's distinction between level-1 and level-2 activities, does philosophy belong to level 2? Is it about concepts rather than the phenomena we use concepts to think about? The answer is that this is a false opposition. This is more obvious in the case of ethics, aesthetics, and political philosophy, because human behaviour is partly guided, and societies and institutions and the arts are therefore partly shaped, by ethical, political, and aesthetic ideas. But it is also true of the philosophy that is concerned with concepts used in natural science. Understanding the phenomena—

 © Mind Association 2017

science, in the broadest sense of the word—is a complex achievement, which depends on a number of different activities: devising theories, testing them experimentally, inventing and making scientific instruments, devising the mathematical and computational techniques which are used to develop theories and interpret experimental data, and inventing the new concepts and understanding the existing concepts in which theoretical ideas are expressed.

In conclusion, art and philosophy are *quite* strange. They are essentially reflective and critical activities, in a way, or to a degree, that the rest of science is not. But I do not believe that neither art or philosophy is quite as strange as Nöe thinks they are, or quite as similar.

The Queen's College, Oxford
john.hyman@queens.ox.ac.uk
JOHN HYMAN

doi:10.1093/mind/fzw060
Advance Access publication 5 January 2017

Judgment and Agency, by Ernest Sosa. Oxford: Oxford University Press, 2015. Pp. 288.

The aim of Ernest Sosa's latest book, which he states in its first paragraph, is to develop his virtue epistemology 'further than before, by taking up issues of metaphysics and ethics (broadly construed) that arise for it'. Part of this development involves clearing up misunderstandings and part of it involves showing how his view can usefully be applied to a wide range of philosophical topics and puzzles. Given Sosa's broad aim, his book covers a lot of ground. The discussions of particular issues will appeal to those interested in those particular topics, and the chapters, on the whole, are self-contained, some being reprints of earlier articles. But, throughout, one finds interconnected and re-emerging themes that are of central importance to all those interested in the nature of belief and knowledge, the nature and extent of epistemic agency, and philosophical methodology. Having Sosa's carefully and forcefully argued views on these topics is most welcome. My discussion will revolve around four of the book's central themes. The first has to do with the importance of carefully making distinctions among related phenomena; it is often the best way to address vexing philosophical problems. Second, that one can and ought to strictly delineate the domain of the epistemic. Third, that what looks like a very demanding conception of knowledge is important, relevant, and indeed central to epistemology, and finally, that Sosa's virtue epistemology has the resources to account for those aspects of our epistemic life that seem to be connected to agency and for which we, as agents, can be held responsible.

That the book is titled 'Judgment and Agency' serves as a clue to the importance of both of these concepts, though one can lose sight of their centrality when led to questions of how Sosa's virtue epistemology can

 © Mind Association 2017

better address traditional philosophical problems concerning lotteries and fake barns. When laying out the summary of chapters in the introduction, Sosa devotes the most space to Chapter 2, 'Virtue Epistemology: Character vs. Competence'. Here is where he considers a challenge to his 'reliabilist' form of virtue epistemology from those who view his theory as quite distinct from what are called 'responsibilist' versions of virtue epistemology, ones which highlight agential virtues, or character traits. It has seemed to many that Sosa's view does not allow a place for such virtues and that a 'virtuous' belief on his account is just one produced by a reliable enough process. Virtuous believers would then be much like virtuous toasters; if a toaster is working reliably enough to succeed in its aim of making toast, it is virtuous; similarly, if a person reliably forms true beliefs then that person is a virtuous believer. Sosa rejects this characterization and tells us that 'reliabilist, competence-based virtue epistemology must be understood broadly, in a more broadly ecumenical way, with responsibilist, agential intellectual virtue at its core' (pp. 3, 36).

To begin to understand how this is so, we can see why this toaster analogy may well apply to a kind of reliabilism, but does not seem to apply to Sosa's *virtue* reliabilism, which has always been concerned with performance normativity. Sosa sees epistemic performances as constituting a species of a larger genus, and thinks that all performances can be evaluated according to how *apt* they are. Sosa opens up the book with a brief reminder that a performance can do well in one or more of three distinct ways: by attaining its aim, by being an exercise of pertinent competence, and finally by being *apt*, which means it reaches its aim through the competence exercised (p. 1). And performances, at minimum, require some agent to perform them. The toaster toasting bread is not a performance. So once we are reminded that knowledge is a successful performance and one in which the believer manifests competence in successfully reaching the aim of belief, namely truth, we should see that the view is not strictly mechanistic in the ways some other reliabilist theories are. Sosa tells us (and offers quotations from his earlier texts to reinforce the point) that mechanisms, as he thinks of them, can cover both simple reflexes as well as higher order complex pondering, and in a footnote says 'dictionaries reveal that a "mechanism" need not reside in a machine. A google search will turn up "trading mechanisms," "defense mechanisms," "mechanisms for dealing with stress," etc.' (p. 37).

If your toaster reliably gives you toast, you will not care *how* it manages to do so. The idea of an object or non-agent *manifesting competence* makes no sense. It is very important to Sosa that one only attains knowledge if the true belief is arrived at 'in the right way'. Throughout the book, Sosa attempts to become clearer and more explicit about what this amounts to. The first chapter, 'The Unity of Action, Perception and Knowledge', argues that one can find this same type of problem about how to offer content to the idea of 'in the right way' emerging for theories of action and perception. What is

 © Mind Association 2017

'the right way' for an intention to be related to the intended act for it to count as a case of acting intentionally? What is 'the right way' for a subject to relate to an object to count as *perceiving* the object? For Sosa, the idea of *manifestation* of competence is key. While he tells us 'What counts as manifestation seems also graspable only in implicit ways, as with etiquette, and not through explicit (and nontrivial) verbal formulation' (p. 26), he does try to provide a deeper understanding by means of analogies and examples. Competences, for Sosa, are types of dispositions. If one is a competent driver, for example, one has a certain skill such that in a range of states and situations one will be successful in driving. Of course, one can only manifest this disposition when driving, and one may succeed in driving without manifesting competence (say if one is intoxicated or tired and could have easily failed). And, ultimately, for an agent to manifest competence in an epistemic performance, it will not be enough for one to possess the disposition to form true beliefs; this disposition will have to account appropriately for the outcome and, at times, this will only be possible if the agent possesses 'responsibilist, agential intellectual virtue'.

Yet it has seemed to some that the whole idea of what Sosa has called 'faulty competence', for example, having a competent memory or a competent faculty of perception, is very different from possessing agential virtues such as open mindedness or intellectual courage and that perhaps the competence of such faculties can be manifest in forming true beliefs without an *agent's* competence, understood as the manifestation of certain intellectual virtues. Sosa insists that, while some knowledge can result from mere faculty competence, he never meant to restrict his view to such knowledge. The mistaken impression, according to Sosa, stems from the nature of many of his examples which involve the competences of simple perception, introspective, or mnemonic knowledge. Sosa says that he would focus on examples 'simple enough to reveal more starkly certain basic problems that any theory of knowledge must solve'(p. 38). And so many of these examples are of what he has called 'animal knowledge' which is simply apt belief, and are meant to distinguish such knowledge (or apt belief) from luckily getting the truth, which is one of those basic problems. When explaining this difference, Sosa often uses the analogy of an archer hitting a target. One's shot is apt if and only if its accuracy manifests its competence. But if we are assessing the performance of a hunter shooting, this evaluation goes beyond whether her shot is apt. Since shot selection is integral to the competence of a good hunter, a hunter 'needs the second-order competence to assess her first order competence and its required conditions', and so even if a hunter's shot is apt '[i]t is reflectively competent if and only if it corresponds to a competent second-order awareness that the shot would be apt' (p. 68). In the realm of epistemic performance, the difference would be between having an apt belief, and forming a reflective judgment about whether one is in a position to aptly believe. For this second order judgment to be apt, one must

 © Mind Association 2017

engage in higher order reflection that manifests competence of intellectual and agential virtues.

Further, Sosa argues that these higher order beliefs, which he calls judgments that result in *judgmental* beliefs about, for example, when to close inquiry or when to endorse lower level credences that arise almost automatically from faculties of memory or perception, are under our volitional control. Thus, he shares what he takes to be Descartes' view that there is a 'distinction between two attitudes that might be called "belief." One is an implicit confidence that suffices to guide our action…The other is an act of judgment made freely and voluntarily, or a disposition to so judge upon considering the relevant question' (p. 238). This distinction between judgmental beliefs and merely 'functional' beliefs which are credences, or degrees of confidence, is one Sosa invokes in many different discussions. He thinks it helps in making sense of what Descartes is up to in the *Meditations*. In particular, he thinks that when Descartes doubts all his 'beliefs' he need not give up the lower level action guiding confidence, but that these can 'do their guiding even if one forbears endorsing them…One can sustain a highly confident credence that p, despite suspending any conscious endorsement of that attitude, and withholding any correlated conscious judgment that p' (p. 239).

Sosa argues that making this distinction between kinds of beliefs can help explain many of our intuitions in the vast Gettier-related examples. For Sosa 'Judgment is affirmation with the intention to *thereby* affirm competently (period) and indeed aptly…Judgment thus involves a second-order stance regarding one's own affirming. When one judges, one affirms with the aim to get it right aptly by affirming' (p. 82, emphasis in original). One can also decide that either positive or negative affirmation is too risky and so suspend judgment and thus 'an aim that might be shared by epistemic affirmation and suspension is the aim *to affirm aptly and only aptly*' (p. 82, emphasis in original) and a judgmental belief again is 'the disposition to judgment upon consideration, when endeavoring to affirm aptly'.

With this is mind, we can now ask why Barney, in fake barn country, fails to know even when he aptly believes there is a real barn in front of him. This is a case where Barney sees a real barn in front of him and believes accordingly, but this is one of the few real barns in a countryside filled with barn façades. Barney falls short of knowledge because, given the danger posed by nearby possibilities, there is some sense in which he has performed unsafely; he could too easily have affirmed incorrectly. Sosa argues that if one wants an account of safety that will explain why Barney falls short, as well as why subjects like Norman the clairvoyant who 'finds [*sic*] themselves believing something, quite reliably in fact, but with no idea how they are doing so, nor even *that* they are doing so' (p. 78, emphasis in original) also falls short, one must locate the unsafety at the level of second order awareness. Barney's *judgment* is apt 'only if he is guided to the aptness of his affirmation (that he does face a barn) by his second-order awareness that if he then affirmed

 © Mind Association 2017

(that he faced a barn) he *would* (likely enough) be right' (p. 79, emphasis in original). If one judges aptly then one's affirmation is safe, but we can see that, for Barney, this is not the case. And so 'Barney and his like fail to judge aptly, and *thus* fail to know (fail to know full well)…[In] order to know full well, Barney must know that if in his conditions he affirmed that he faces a barn, not easily would he be wrong' (p. 79, emphasis in original).

Still, Barney and his like may know *in some sense*. According to Sosa, when it comes to cases of the type involving Norman the clairvoyant there is a close to even split among epistemologists about whether such subjects know. He introduces a case of this type taken from ordinary life by asking us to consider what goes on when we take an eye exam. As I get to the bottom rows, I start to lose confidence that I am getting it right and am in effect 'guessing' whether, say, it is an 'E' or an 'F'. 'Suppose however it turns out, that (unbeknownst to me) I am in fact getting it right year after year at a line where I am thus unsure' (p. 74). In such a case I am affirming and doing so in the endeavour to get it right and I am manifesting competence in doing so. How then do we assess my performance? Sosa says 'We are here conflicted. *Somehow* I *do* know what letters I see, as shown by my impressive reliability. But there is also a pull to say I do not *really* know' (p. 75, emphasis in original). Sosa explains this by saying while we may know at some very basic level, at the point where we start guessing, we are no longer even aiming for aptness of the affirmation; we are just aiming to do what the test requires. We are still aiming at truth, trying to get it right, but aiming to affirm correctly and aiming to affirm aptly differ. The latter requires reflective competence and apt judgment.

Those familiar with Sosa's epistemology will be accustomed to the idea that there are levels of knowledge, corresponding to iterated aptness. We have seen the difference between animal and reflective knowledge, but we see now that the idea of *full aptness*, where one's reflective knowledge guides one's judgments, relates to an even higher level. One can aptly affirm that p, but one can also aptly affirm the aptness of one's affirmation that p. This is, as we have seen, what Sosa calls knowing full well. In his book of that title this even higher level was introduced, but it wasn't entirely clear whether the difference between a judgment being sufficiently apt to count as reflective knowledge and it being *fully apt*, which it can be only if 'it is guided to aptness through the agent's reflectively apt risk assessment' (p. 69) was an important difference for epistemology.

Now that Sosa is focusing attention on reflective knowledge, he attempts to show that this kind of knowledge is actually crucial in trying to explain the many intuitions about knowledge possession that have been at the centre of traditional epistemology. He says 'The more important concept, it now seems clear, is that of knowing full well' (p. 74). It is important for Sosa that questions about epistemic agency are not taken as orthogonal to the traditional problems of the nature and value of knowledge. It is by recognizing that some

 © Mind Association 2017

knowledge is only attainable if one *chooses* to endorse what seems true while endorsing also that this choice is apt that one can see why in certain cases people fall short even when it seems they have manifested competence to some degree and end up with true beliefs because of this competence.

Many recent discussions of epistemic or doxastic agency argue that the epistemic and ethical realms are not as distinct as traditional epistemology has assumed. As we have seen, those who have been critical of competence-based, reliabilist virtue epistemology think that it is important to recognize that certain moral virtues have a role to play in virtuous believing. Further, once one accepts that there is a volitional component in what we believe, it seems the norms of evaluation will be much more continuous with those which govern action. But it is important for Sosa that this unity not be overstated and that epistemic performances remain distinct from performances of other kinds. It is significant that he tells us he is concerned with discussing issues of ethics 'broadly construed' that arise for his theory, since he thinks that certain kinds of ethical issues are inappropriately applied in the context of epistemology. If I evaluate you as being callous or aggressive in not allowing your 10 year old to take back a chess move, I am not thereby assessing anything about you *qua* chess player. Sosa argues that epistemic performances can be insulated in a similar way your performance in chess.

So Sosa wants to make room for 'responsibilist, agential intellectual virtue' as being essential for the highest degree of epistemic competence, but also wants to exclude moral or ethical virtues from the epistemic domain. To make this distinction he says 'we should distinguish between an intellectual ethics that is purely epistemic, and an intellectual ethics that is properly a part of ethics' (p. 43). While some ethical virtues and competences such as open mindedness and intellectual courage can be applied in the epistemic domain '[t]he study of such ethical intellectual competences and virtues would be part of applied ethics…Business ethics is a branch of ethics that studies ethical issues concerning the practice of business in particular. A correlate intellectual ethics is thus a branch of ethics that studies ethical issues concerning scientific or other research, and concerning the value of various sorts of knowledge for human flourishing, and concerning issues of 'the acquisition and retention and sharing of such knowledge. Et cetera' (p. 44). To determine what counts as a purely epistemic virtue we need to take the question as given and focus on how best to settle the question, by affirmation, denial or suspension. The kinds of virtues that are relevant here are ones which will render such a choice apt. In making this choice 'one should exercise proper *care* and *attentiveness*' and these can be 'stable character traits of an epistemic agent' (pp. 44-5, emphasis in original). Sosa also mentions some virtues of inquiry such as those relevant to 'putting one in a position to know' and 'intellectual perseverance' that, while not knowledge-constituting competences, are all '*purely intellectual* virtues' that are 'integral to a *purely epistemic* intellectual ethics' (p. 45, emphasis in original).

 © Mind Association 2017

Nevertheless, given that the focus of this book is on reflective, conscious judgmental beliefs, Sosa says his main interest is on these *acts:* 'Our emphasis in on humans as social animals who reason, often consciously, who do so individually and collectively, both practically and theoretically. So we focus on the practical and theoretical acts by means of which humans accomplish all of that…[If] our interest is rational actions or beliefs that are outcomes of reasoning…then acts are where the action is' (pp. 67-8). Focusing on acts, but acts to which ethical evaluation does not apply, requires making sense of *purely* epistemic or intellectual acts. Sosa says more about such epistemic purity when considering the issue of pragmatic encroachment and what counts as a competence being 'reliable enough' to guide our judgments. He asks 'If we put aside pragmatic concerns such as whether the check will bounce, or whether we will be late for a meeting in another city, what then determines whether a competence is epistemically reliable *enough*?'(p. 172). If we are concerned only with the 'purely epistemic', the only relevant competence is the disposition 'to discern the true from the false in a particular domain' (p. 172, emphasis in original). How reliable is 'reliable enough' to count as competent? This is a difficult question, and Sosa devotes some time to addressing it, but it is important for him that, when addressing it, we restrict ourselves to the 'domain-internal standards' of epistemic success, just as we appeal to domain-internal standards in tennis or hunting to determine competence in those domains: 'Whether a particular performance is appropriate within either domain depends on how appropriately that performance is meant to contribute – and how appropriately it does contribute – to a pattern of activities with enough probability of attaining domain-internal success, such as that of the hunt or that of the match' (p. 173).

I will close my discussion with some questions and concerns. We have seen that Sosa thinks we can solve or dissolve many philosophical problems by distinguishing between certain kinds of belief and certain levels of knowledge. I worry that it can be difficult, when we are concerned with a creature with the capacity for higher order reflection, to usefully maintain the distinction between these two kinds of 'belief'. When Sosa is defending his view against an overly mechanistic interpretation, and insisting that 'responsibilist competences' have always been present, he quotes a passage from the earliest presentation of his virtue reliabilism: 'The beliefs of a *rational* animal hence would seem never to issue from *unaided* introspection, memory, or perception. For reason is always at least a silent partner on the watch for relevant data, a silent partner whose very *silence* is a contributing cause of the belief outcome' (p. 37, emphasis in original). What this picture suggests is that all our beliefs, even the so-called merely 'functional ones', are being tacitly endorsed, and thus seem to be products of our agency just as our more consciously expressed judgments are. One may well think that what transforms a mere 'seeming' into a 'belief' is that it can be rationally evaluated and the person concerned can be held responsible for maintaining it.

 © Mind Association 2017

But, throughout much of his discussion, Sosa argues that it is only at the higher, reflective conscious state that we locate volition and responsibility.

Consider again Sosa's interpretation of Descartes. Sosa says that to put beliefs in doubt, on this view, 'is *not* to disable them from functioning in the guidance of action, nor even in the reasoning required for inquiry. Compatibly with a belief's retention of its animal/*cognito* status and its ability to provide the guidance expected of such beliefs, the believer might refuse to endorse his belief when it is brought to consciousness for rational inspection' (p. 239, emphasis in original). The main difference, for Sosa, between Descartes and the Pyrrhonists, is that the Pyrrhonists never so endorse while, for Descartes, some beliefs can be rationally endorsed. But how can one retain a belief and not endorse it? Usually, once one recognizes that a belief is not worthy of endorsement, one thereby loses the belief. And those beliefs that are resilient, such as prejudices learned in childhood that one does not endorse but cannot quite shake, are problematic. Also, given the role of reason as the 'silent partner', it is hard to see how any seeming or credence that rises to the level of belief is not endorsed.

The idea of 'epistemic purity' which seeks to carve out an autonomous epistemic domain along the lines of tennis or hunting is problematic. The first thing to note is that Sosa's cited 'historical antecedents', namely Sextus and Descartes, did not think of epistemology along such lines. For Descartes, errors in belief and in action are both failures of the will; they are both moral errors. Sosa himself points to an important difference between the epistemic domain and others in Chapter 8, 'Social Roots of Human Knowledge'. After emphasizing the need to carve out a pure epistemic domain, he says that trying to define 'the epistemically successful life' is not as simple as defining success in tennis or hunting, and this is partly because '[as] humans and as fellow members of our communities and of our species, we depend crucially and variously on the acquisition and sharing of information' (p. 173). And he says that in thinking about what counts as being epistemically competent we need to factor in these collective and practical concerns: 'Here non-epistemic factors do plausibly bear' (p. 173). But once it is recognized that epistemic performances and their success cannot be isolated from the wider needs and concerns of the community, the idea of their purity is put into question and we can see that they are importantly different from performances in tennis or hunting.

There is much more in this book than I have been able to discuss here. Sosa devotes quite a lot of space to explaining what he takes to be the relevant safety condition on knowledge, distinguishing it from sensitivity conditions and other safety conditions (Chapters 4 and 5), and says much more about the nature of competence (Chapters 4, 8, and 9)! His discussion about the various kinds of philosophical explorations and the importance of recognizing them is extremely illuminating (Chapter 1). I have said next to nothing about his view of intentional action (Chapter 7) or his take on how to

 © Mind Association 2017

interpret Sextus and the possibility of living without beliefs (Chapter 10). Those interested in deepening their understanding of these topics and Sosa's view on them will benefit a great deal from these discussions.

University of Richmond
mmccorm2@richmond.edu
MIRIAM SCHLEIFER MCCORMICK

doi:10.1093/mind/fzw058
Advance Access publication 7 January 2017

© Mind Association 2017

Erratum

A book review of "Believing Against the Evidence" by Miriam Schleifer McCormick (Mind 125–499, 942–945) was credited to the wrong author. The Book Review was incorrectly credited to:

José Luis Bermúdez
Texas A&M University
jbermudez@tamu.edu

It should have been credited to:

Ram Neta,
University of North Carolina at Chapel Hill
Neta@email.unc.edu

This has now been corrected in the online version of the book review.

The publisher apologises for this error.

Mind, Vol. 126 . 501 . January 2017 © Mind Association 2017
doi:10.1093/mind/fzw075 Advance Access publication 6 January 2017

Quality & speed

Advance Access publication from Oxford Journals

Articles published online ahead of print are available to read and cite with Advance Access.

www.oxfordjournals.org

OXFORD
UNIVERSITY PRESS

Russellian Acquaintance and Frege's Puzzle

Donovan Wishon
University of Mississippi
dwishon@olemiss.edu

In this paper, I argue that a number of recent Russell interpreters, including Evans, Davidson, Campbell, and Proops, mistakenly attribute to Russell what I call 'the received view of acquaintance': the view that acquaintance safeguards us from misidentifying the objects of our acquaintance. I contend that Russell's discussions of phenomenal continua cases show that he does *not* accept the received view of acquaintance. I also show that the possibility of misidentifying the objects of acquaintance should be unsurprising given underappreciated aspects of Russell's overall theory of knowledge and acquaintance. Finally, I consider the radical impact that Russell's actual views on acquaintance have for our understanding of his well-known George IV case in 'On Denoting'. In particular, I argue that Russell's treatment of the George IV case is not a one-size-fits-all solution to Frege's Puzzle and provides no support for the received view of acquaintance.

1. Introduction

Russell's Theory of Descriptions is at best a partial solution to Frege's Puzzle. Frege's Puzzle is the problem of explaining how it is that distinct co-referential expressions (and their analogues in thought) can make different contributions to the cognitive significance of the utterances or thoughts in which they occur.[1] Following Frege, this problem is most often discussed in the context of identity statements: an identity sign flanked by two referring expressions. In these paradigmatic cases, the problem is to explain how thoughts or utterances of the form '$a = a$' and '$a = b$' can differ in cognitive significance if 'a' and 'b' co-refer. In such cases, the first is analytic and a priori while the second need not be. But it is important to keep in mind that similar problems arise whenever we use (rather than mention) distinct expressions (or concepts) whose semantic contributions at the level of

[1] In saying here and elsewhere in this paper that linguistic expressions have 'analogues in thought', I mean to be neutral on the issue of whether there is a genuine, full-blown 'language of thought' as well as on the issue of whether Russell thinks that there is one.

Mind, Vol. 126 . 502 . April 2017
doi:10.1093/mind/fzv193 Advance Access publication 13 October 2016

© Wishon 2016

reference or denotation is the same and yet whose contributions to cognitive significance differ. For instance, neither 'Hesperus is visible in the evening sky' nor 'Phosphorus is visible in the evening sky' seem analytic or a priori, and yet there is a clear difference in cognitive significance between them despite the fact that 'Hesperus' and 'Phosphorus' are co-referring expressions.[2]

Russell's Theory of Descriptions (when combined with his well-known Descriptive Theory of Ordinary Names) provides two powerful resources for solving Frege's Puzzle. First, it draws a distinction between the surface grammatical form of an utterance and its underlying logical form, allowing Russell to claim that most thoughts or utterances that *appear* to have the logical form '$a = b$', for example, do not *really* have it. And second, it maintains that most expressions in natural language which *appear* to be referring expressions instead function logically as if they were truncated descriptive phrases (or 'incomplete symbols'), ones whose contributions to the actual logical form of the thoughts or utterances in which they occur are (restricted) existential quantifiers which are determined in context.

As a consequence, thoughts and utterances of the *grammatical* form '$a = a$' either assert the self-identity of a particular object *a* or that there is a unique satisfier of such and such description which is the unique satisfier of that very description, both of which are trivial (aside from their existential import). However, most thoughts and utterances of the *grammatical* form '$a = b$' either assert that a particular object *a* uniquely satisfies such and such description or that there is an object which uniquely satisfies two different descriptions, both of which are often substantive (in addition to their existential import).

In this way, Russell's Theory of Descriptions (combined with the Descriptive Theory of Ordinary Names) promises to solve Frege's Puzzle for a significant class of thoughts or utterances. But it cannot by itself solve Frege's Puzzle for thoughts or utterances of *all* logical forms. It cannot explain the difference in cognitive significance, if any, between thoughts or utterances with the *logical* form '$a = a$' and the *logical* form '$a = b$' where '*a*' and '*b*' are co-designating, genuinely singular, referential devices. Nor can it explain the difference in cognitive significance, if any, between thoughts or utterances of the form 'the unique F is the unique F' and the form 'the unique F is the unique

[2] Also consider familiar substitution problems in attitude ascriptions, as well as cases such as 'Hesperus is visible in the evening sky only if Hesperus is' and 'Hesperus is visible in the evening sky only if Phosphorus is'.

 © Wishon 2016

G' where 'the unique F' and 'the unique G' contribute the same elements and (restricted) existential quantifiers, in the same contexts, to thoughts and utterances in which they occur. (Think of 'the capital of *this* country' and 'the capital of *that* country' where the '*this*' and '*that*' co-refer.)[3]

According to a cluster of widely-held and closely-related views, which I'll together refer to as 'the received view', Russell has a simple, elegant solution to both lingering problems. The first part of this solution is to place a substantive constraint, which is often called 'Russell's Principle', on what it takes to think or talk significantly about things. 'Russell's Principle' demands that we know *which* things we are thinking or talking about if we are to think or speak with understanding. Russell enunciates 'Russell's Principle' clearly in *The Problems of Philosophy* (hereafter *POP*) when he claims that:

> It is scarcely conceivable that we can make a judgement or entertain a supposition without knowing what it is that we are judging or supposing about. We must attach *some* meaning to the words we use, if we are to speak significantly and not utter mere noise; and the meaning we attach to our words must be something with which we are acquainted. (p. 91, emphasis in original)

However, it is important to recognize that 'Russell's Principle' does not all by itself solve the lingering versions of Frege's Puzzle. After all, one can presumably know that 'Cicero' refers to the particular individual Cicero with whom we have ordinary (non-Russellian) 'acquaintance' and that 'Tully' refers to the particular individual Tully with whom we have ordinary (non-Russellian) 'acquaintance' without knowing that Cicero *is* Tully.[4]

The second part of Russell's supposed solution to Frege's Puzzle is to impose a particularly demanding version of 'Russell's Principle' on our thought and talk about things: we can only think or talk about things with understanding if we have an epistemically robust kind of direct acquaintance with them, or if we think of them indirectly as the unique satisfier of some description composed entirely out of elements with which we have such acquaintance.[5] In particular, the received

[3] During the period under discussion, Russell takes 'this' and 'that' to be paradigm cases of referring expressions.

[4] Also consider Kripke's (1979) cases of 'London' and '*Londres*' or 'Paderewski' and 'Paderewski'.

[5] From now on I will use the term 'acquaintance' in Russell's technical sense, rather than in the ordinary sense.

 © Wishon 2016

view has it that Russell's technical notion of acquaintance, together with his restrictions on objects we can be acquainted with, *precludes* cases in which *a* and *b* are the same, we are acquainted with the thing they both are, and yet we do not know the right answer when we consider whether '*a* is *b*' using two distinct co-referring expressions '*a*' and '*b*', each of whose meaning we understand.

On this reading of Russell, cases such as that of Cicero and Tully above simply cannot arise when bona fide Russellian acquaintance is involved. Nor can cases of the form 'the unique F is the unique G' where we are acquainted in Russell's technical sense with all the elements that the two descriptive phrases contribute to the logical form of our thought and talk in a particular context. Thus, Russell's Theory of Descriptions (and the Descriptive Theory of Ordinary Names), when supplemented with the received view of Russellian acquaintance, constitutes a tidy and alluring possible solution to Frege's Puzzle.

Nevertheless, there is little evidence that Russell himself holds the received view of acquaintance, and in my estimation, none of the evidence is compelling. On the contrary, Russell recognizes and discusses cases throughout his work in which 'Russell's Principle', under the strong interpretations required by the received view, is false. Most importantly, Russell notes on several occasions that there are phenomenal continua cases in which we are acquainted with both *a* and *b* without thereby knowing whether we are acquainted with two distinct things, or one thing twice over. In other words, *a* and *b* can be the same thing, we can be acquainted with the thing they both are, and yet *not* know the truth-value of '*a* is *b*' despite knowing the meanings of the co-referring expressions involved. But if this is so, then Russell does not solve Frege's Puzzle by supplementing his Theory of Descriptions (and Descriptive Theory of Ordinary Names) with 'Russell's Principle' strongly construed. In my view, the evidence suggests that he never intended acquaintance to help solve Frege's Puzzle in the first place.

In this paper, I proceed as follows. First, I show that a number of Russell interpreters, such as Gareth Evans, Donald Davidson, John Campbell, and Ian Proops, have indeed attributed to him various versions of the received view of acquaintance. Second, I present his discussions of phenomenal continua cases, which show that he recognizes that 'Russell's Principle', strongly interpreted, is false. Third, I explore what Russell actually says about the nature of acquaintance and its role in knowledge and reference, while drawing attention to the distinction Russell makes between the *ordinary* significance and *logical*

 © Wishon 2016

significance of our thought and talk. Lastly, I address his discussion of George IV in 'On Denoting' (hereafter OD) and elsewhere, which has (wrongly) convinced many interpreters that Russell does intend the combination of acquaintance and the Theory of Descriptions to solve Frege's Puzzle wholesale.

2. The received interpretation

Many previous interpreters of Russell have attributed to him some version or other of what I call the received view of acquaintance, at least during some period of his thought. In its most general form, the received view of acquaintance holds that when a subject is acquainted with *a* and *b*, designates *a* and *b* with distinct genuine proper names, and uses these genuine proper names with understanding to consider whether *a* is *b*, he or she cannot fail to determine the correct answer if the question arises. But it also encompasses the even stronger view that when a subject is acquainted with *a* and *b*, he or she *thereby* knows whether or not *a* is *b*. In either case, the central idea is that acquaintance is supposed to safeguard us from misidentifying the objects of our acquaintance.

One of the most influential of these interpretations is Gareth Evans' in his *The Varieties of Reference* (1983). According to Evans, Russell holds the received view of acquaintance partly because of his deep-rooted Cartesian sentiments and partly because he thinks it is the only satisfactory way of complying with 'Russell's Principle'. Indeed, Evans claims that Russell requires that a subject have what he calls a 'discriminating conception' of a thing in order to be able to think or talk about it with understanding, where a *discriminating conception* is a conception of a thing that enables the subject to distinguish it from all other things (p. 65). Evans maintains that Russell sees only two ways of attaining such a demanding discriminating conception of a thing: either by being acquainted with it or by thinking of it as the unique satisfier of a description composed from elements all of which the subject is acquainted with. As Evans puts it:

> [Russell] held a theory of thought which incorporated the principle that it is not possible for a person to have a thought about something unless he knows which particular individual in the world he is thinking about. Russell also held that there are only two ways of discharging this requirement: one is to be, or to have a memory of being, 'acquainted' with the object concerned, and the other is to think of the object as the unique satisfier of some description. (Evans 1983, p. 44)

 © Wishon 2016

Regarding acquaintance itself, Evans correctly notes that Russell sees it as a two-place relation that either holds between a subject and an object, or not at all. However, Evans also recognizes that even if acquaintance is a two-place relation between a subject and an object, this alone does not secure a discriminating conception of the object and thereby preclude the possibility of misidentifying it. For a subject could conceivably stand in this two-place relation to one and the same thing many times over in different circumstances or on different occasions without realizing that it is the self-same object. Evans argues that this is the principal reason why Russell radically restricts our acquaintance to sense-data and other 'Russellian objects'.[6] For these are the sorts of objects, one might plausibly think, that a subject presumably could not fail to identify or re-identify in different circumstances. Indeed, he says:

> Russell himself had an excuse for his espousal (in effect) of the ordered-couple conception of monadic Russellian thoughts, in that he restricted the objects of such thinking mainly to items which were conceived to be so fleeting and insubstantial that it seemed unintelligible to suppose a person might identify the same one twice without knowing it was the same. But this excuse is not available to those who extend the ordered-couple conception of Russellian thoughts outside the sphere of thoughts about objects of Russellian acquaintance. (Evans 1983, p. 82)

Hence, to summarize Evans' take on Russell, the special nature of the objects of acquaintance, together with the relational character of acquaintance, is what provides a subject with a discriminating conception of an object sufficiently demanding to satisfy 'Russell's Principle'.

Evans is not the only influential philosopher to misinterpret Russell along these lines. In his 'What is Present to the Mind?' (1989), Donald Davidson attributes a very similar view about the nature of acquaintance, and of 'Russellian objects', to Russell. Davidson disputes the general view that whenever we think about a thing, we have special objects before our mind (whether they are sense-data or propositions) which are impossible to misidentify and which fix the contents of our thought (p. 53). Furthermore, he aims to show that there is no *special* epistemic relation, such as Russellian acquaintance, which a subject bears to objects that enables him or her to entertain genuine *de re* thoughts about them. Rather, he contends, '*any* property of an object may, under suitable conditions, be considered the relevant identifier' which allows the subject to think *de re* thoughts about it (p.56).

[6] The term 'Russellian objects' comes from Proops (2011).

 © Wishon 2016

On Davidson's view, the problem with Russell's view is that it is committed to a thoroughly Cartesian conception of thought, one according to which we cannot be wrong about what it is that we are thinking. He argues:

> What lies behind some of these attempts to characterize the special relation between the mind and its objects is, of course, the Cartesian drive to identify a sort of knowledge which is guaranteed against failure. If this search is combined with the assumption that all knowledge consists in the mind being in psychological contact with an object, then objects must be found about which error is impossible—objects that must be what they seem and seem what they are. (Davidson 1989, p. 56)

Moreover, Davidson (like Evans) takes it that Russell's solution consists not only in taking thought to be ultimately grounded in a special epistemic relation subjects bear to things, but also in restricting the things we can think directly about to special 'Russellian objects'. But, Davidson argues, not only is there no such thing as a single, special kind of epistemic relation that allows us to think *de re* thoughts about things:

> there simply are no such ['Russellian objects']. Not even appearances are everything we think they are. Nor can the 'aspects' of sense data, if they really are objects, be protected from one or another sort of misidentification. (Davidson 1989, p. 56)

Consequently, he remarks:

> If the mind can think only by getting into the right relation to some object which it can for certain distinguish from all others, then thought is impossible. If a mind can know what it thinks only by flawlessly identifying the objects before it, then we must very often not know what we think. (Davidson 1989, p. 56)

Incidentally, we will see that Russell himself would very much agree with Davidson on this point. Still, Davidson clearly *takes* Russell to be committed to the received view of acquaintance.

More recently, in his 2009 'Consciousness and Reference', John Campbell too attributes the received view of acquaintance to Russell. On Campbell's view, one of the central issues motivating Russell's use of the notion of acquaintance is the important connection he sees between our consciousness of things and our knowledge of singular, demonstrative reference. He recognizes that for Russell, acquaintance is a non-propositional species of direct epistemic contact that a subject can have with things that both ground the reference of simple referring expressions and provide him or her with knowledge

 © Wishon 2016

of their reference (or better, I think, at least positions the subject to acquire such knowledge of reference). Campbell also rightly sees Russellian acquaintance with things as psychologically and epistemologically more primitive than any propositional knowledge of truths about them, or descriptive designation of them by means of such truths.

In trying to make sense of such primitive, non-propositional knowledge by acquaintance, Campbell appeals to G. E. Moore's (1903) notion of the 'transparency of experience'. The basic idea here is that experience is a generic two-place relation between a subject and the objects of experience such that by specifying the objects experienced, we have thereby fully specified the character of the experience. The problem with this view, Campbell acknowledges, is that it makes it difficult to account for manifest sameness of reference across changes in the character of experience. According to Campbell, Russell was sensitive to this problem and responded to it by simply restricting the objects of experience to those that can only be experienced in one way. In other words, Campbell argues, Russell's strategy was to:

> keep the idea that we should characterize consciousness in Moore's terms, as the holding of a generic relation between the self and an object. But we should restrict the range of the relation. We should keep it restricted to objects of which the subject automatically has *comprehensive* knowledge. (Campbell 2009, p. 656, emphasis added)[7]

However, Campbell warns against our taking Russell's view to be that acquaintance exhaustively *reveals* the essential nature of the objects of experience, despite Russell's well-known remarks in *POP* that a subject knows a colour 'perfectly and completely' when he or she is acquainted with it (p. 73). For as Campbell notes, this suggests that acquaintance is a species of knowledge of truths about essences rather than a species of knowledge more primitive than, and wholly distinct from, *any* knowledge of truths. Indeed, he contends:

> Russell's remark about completeness of knowledge should not be read as relating to propositional knowledge of essences; the remark is, rather, his response to the problem of partial awareness. The idea is that the colours are such that there is nothing partial about our awareness of them; so we can characterize acquaintance with them fully merely by saying which colours are being encountered. There are not, on Russell's view, different

[7] Campbell's solution is to see acquaintance as a three-place relation between a subject, a standpoint, and an object.

 © Wishon 2016

ways of being acquainted with one and the same colour. The point is rather that the knowledge of the thing is complete; there is no further, non-propositional knowledge of the thing to be had, once you have encountered it in experience. When put generally, this leads to the doctrine of special objects of awareness, which can be encountered in only one way. (Campbell 2009, p. 661)

Finally, Proops (2011) attributes to the early Russell (prior to 1914) roughly the same view about the nature of acquaintance with objects as does Campbell. According to Proops, reflection on the case of George IV in OD made Russell very much aware that his Theory of Descriptions is at best a partial solution to Frege's Puzzle, and to substitution problems in general (p. 156).[8] Russell's solution prior to his 1914 *Our Knowledge of the External World* (hereafter *OKEW*), Proops argues, was to supplement the Theory of Descriptions with the view that acquaintance fully discloses to a subject the objects of his or her acquaintance. Proops glosses the relevant notion of 'full disclosure' with the following epistemological principle:

> Full Disclosure: Whenever a subject, *S*, is acquainted (in Russell's technical sense of that term) with an object, *x*, *S* is acquainted with every part of *x*. (Proops 2011, p. 152)

He takes it that this is precisely the sense in which Russell takes our knowledge of a thing to be 'perfect' and 'complete' when we are acquainted with it: we are acquainted with every part of it. What this ensures is that when a subject is acquainted with a thing, it has no hidden parts and hence there is no possibility of the subject having different, partial 'takes' on it. Consequently, Proops suggests, the subject's acquaintance with a thing would preclude the possibility of misidentifying it, thereby blocking the lingering threat of Frege's Puzzle.

Proops bases his interpretation on a passage in *OKEW* in which Russell maintains that acquaintance does not come in degrees, but rather is 'either complete or non-existent' (pp. 144–5). One prima facie problem with Proops' interpretation is that Russell claims in the same passage that we cannot literally become 'better acquainted' with any object but only 'with more parts of a certain whole'. But if acquaintance satisfies Full Disclosure, then there is no genuine sense

[8] Proops never explicitly claims that Russell is concerned with Frege's Puzzle. Instead, Proops plausibly takes Russell to want to defend what Proops calls the 'Saving (truth-) Value Principle' (or 'SV'), which he formulates as follows: 'The substitution of identical propositional constituents within a proposition preserves that proposition's truth-value'. Still, Proops *does* see Russell as concerned with substitutions of propositional constituents in attitude ascriptions and to that extent does see him as concerned with Frege's Puzzle.

 © Wishon 2016

to be made of becoming 'better acquainted' with a thing since we would already be acquainted with every part of it. Proops anticipates this worry and insists that it is not a problem 'since Full Disclosure is a substantive thesis, [and] the idea of becoming better acquainted with something is not unintelligible' (p. 145).[9] In §4, I will offer a more natural reading of Russell's remarks.

In any case, it is doubtful that Russell holds the received view of acquaintance. Though Russell clearly is interested in *some* substitution problems when discussing George IV in OD and elsewhere, it is doubtful that his treatment of that case is meant to be a *general* solution to substitution problems, and in any case it offers little support for 'Full Disclosure' or for the received view of acquaintance. In point of fact, there is a great deal of evidence that Russell's *actual* views concerning acquaintance are primarily motivated by very different concerns than answering Frege's Puzzle, contrary to prevailing Russell interpretation.[10] But before I advance my own interpretation of his views on the nature of acquaintance, I will show that Russell indeed recognizes that 'Russell's Principle', strongly interpreted, is false.

3. Russell on phenomenal continua cases

There are a number of aspects of Russell's overall theory of knowledge which suggest that he did not in fact endorse the received view of acquaintance, or 'Russell's Principle' strongly construed. But the most decisive evidence for this comes from his frequent discussions of so-called 'phenomenal continua cases', and related phenomena.[11] Phenomenal

[9] One additional worry with Proops' interpretation is that Russell (1911 and 1992) explicitly notes the possibility of being acquainted with complexes or facts without being acquainted with all their constituents. I would like to thank Russell Wahl for drawing attention to this second point. See Clark (1981) for similar remarks.

[10] A far from exhaustive sampling of recent philosophers who seemingly attribute some version of the received view of acquaintance to Russell includes Austin (1990), Bach (2010), Bigelow and Pargetter (2006), Jeshion (2010a; 2010b), Johnston (1992), Kaplan (2005), Neale (2005), Recanati (2010), Richard (1997), Salmon (1986), Soames (2010), Sullivan (2010), Szabó (2005), White (2007), and Williamson (2000).

[11] For example, see *POP* (pp. 215–6), *OKEW* (pp. 141–9), and Russell (1923, pp. 148–9; 1927, p. 228; 1940, pp. 98–101; and 1948, pp. 276–8, 411–2). One might question whether Russell's post-1921 discussions of phenomenal continua are relevant to the received view of acquaintance since it is widely supposed that Russell abandons his notion of acquaintance in *The Analysis of Mind* (hereafter *AMi*) (1921, pp. 142–4). I think this is another point on which received interpretations of Russell are mistaken (and not simply because the change came in Russell 1919). In Wishon (2015),

 © Wishon 2016

continua are cases in which one item or quality transitions into another item or quality such that no difference can be detected between two distinct points *a* and *b* in the transition, nor between two distinct points *b* and *c*, and yet some difference can be detected between points *a* and *c*. Such phenomenal continua cases conflict with the received view of acquaintance in that they generate circumstances in which a subject can evidently be acquainted with *a* and *b*, designate *a* and *b* with distinct referring expressions he or she understands, consider whether or not *a* is *b*, and yet fail to know what the right answer is.

Though Russell discusses phenomenal continua cases at least as early as chapter thirteen of *POP*, his earliest sustained discussion of them appears in his 1914 Lowell Lectures, which were published the same year as *OKEW*. In these lectures, Russell's principal aim is to argue that we can usefully deploy logico-mathematical methodologies to solve an assortment of traditional philosophical problems. For instance, Russell argues in chapter five, 'The Theory of Continuity', that we can usefully deploy what he elsewhere calls 'logical analysis' to construct models exhibiting compactness to solve Zeno's Paradox (pp. 129–52).

However, this is not Russell's only aim. He also wants to answer what he takes to be one of the most powerful objections to his use of 'logical analysis' to solve philosophical problems, namely, that we cannot do so because experience shows us that the world is a single unity which does not have the parts, or discrete units, necessary for veridical logico-mathematical modelling, and in particular shows the mathematical continuum 'to be inadequate to the facts of sense' (*OKEW*, pp. 129–52). Following Russell's lead, we might call this

---

I argue that rather than giving up the notion of acquaintance entirely, Russell simply changes his view about its nature. In particular, he moves away from his earlier view that acquaintance is a dyadic relation between two distinct relata: a mental subject and an object which may or may not be mental. In its place, he adopts the view characteristic of his neutral monist phase that there is neither a fundamental mental subject nor a distinction between the act of sensation and its object. Nevertheless, he continues to see acquaintance as of great philosophical significance, using it for instance as a key part of his argument in *Human Knowledge: Its Scope and Limits* (hereafter *HK*) that the mind-body problem is an epistemic one rather than an ontological one (pp. 224, 240–7). Furthermore, in his 1956 *My Philosophical Development* (hereafter *MPD*), he says that 'such words as "awareness", "acquaintance", and "experience" had to be *redefined*' (p. 136, emphasis added) and that 'I have maintained a principle, *which still seems to me completely valid*, to the effect that, if we can understand what a sentence means, it must be composed entirely of words denoting things with which we are acquainted or definable in terms of such words' (p. 169, emphasis added).

 © Wishon 2016

British Monistic Idealist challenge to logical analysis 'the Thesis of Interpenetration', which he presents as follows:

> It is often urged that, as a matter of immediate experience, the sensible flux is devoid of divisions, and is falsified by the dissections of the intellect. Now I have no wish to argue that this view is *contrary* to immediate experience: I wish only to maintain that it is essentially incapable of being *proved* by immediate experience. (*OKEW*, pp. 143–4, emphasis in original)

The argument against logical analysis here appears to be something along the following lines. Firstly, our experiential field does not present itself to us *as being* of a manifold of discrete sense-data. Secondly, the best, and seemingly only, explanation for this fact is that our experiential field is *not* of a manifold of discrete sense-data. And, thirdly, Russell's 'analytic' methodology is only adequate for analysing discrete units ('logical atoms'). Therefore, Russell's analytic method cannot be used to solve traditional philosophical problems.

In response to this challenge, Russell's strategy is to let the first and third premises pass but to reject the second one. The problem with the second premise, he argues, is this: although the view that the world as we experience it is 'devoid of divisions' is empirically adequate, so is his own sense-data view (*OKEW*, p. 144). For, he points out, if there are phenomenal continua cases, as we have good reason to believe, then they would present themselves to us in way that is indistinguishable *for us* from an experiential field 'devoid of divisions'. So, he continues, 'the question whether sense-data are composed of mutually external units is not one which can be decided by empirical evidence' (*OKEW*, p. 143).[12] Consequently, the British Monistic Idealist's argument is unsupported and gives us no reason to abandon his analytic methodology.

In making his case, Russell relies on two different kinds of phenomenal continua cases: a diachronic case of our experience of weight and a synchronic case of our experience of shades of colour. A diachronic phenomenal continua case is one in which a at time t_1 is indistinguishable from b at time t_2 and b at time t_2 is indistinguishable from c at time t_3 but a at time t_1 is distinguishable from c at time t_3. A synchronic case is one in which a, b, and c are all co-present, where a and b or b and c cannot be distinguished pair-wise, but a and c can be. In describing the first, diachronic case, Russell says:

[12] These remarks also provide strong evidence that Russell did not conceive of acquaintance as *revealing* the essential nature of its objects, as some interpreters have maintained.

 © Wishon 2016

> When we are considering the actual data of sensation in this connection, it is important to realise that two sense-data may be, and *must* sometimes be, really different when we cannot perceive any difference between them. An old but conclusive reason for believing this was emphasised by Poincaré. In all cases of sense-data capable of gradual change, we may find one sense-datum indistinguishable from another, and that other indistinguishable from a third, while yet the first and third are quite easily distinguishable. Suppose, for example, a person with his eyes shut is holding a weight in his hand, and someone noiselessly adds a small extra weight. If the extra weight is small enough, no difference will be perceived in the sensation. After a time, another small extra weight may be added, and still no change would be quite easily perceptible; but if both weights had been added at once, it may be that the change would be quite easily perceptible. (*OKEW*, pp. 141–2, emphasis in original)

And in describing the second, synchronic case, he remarks:

> Or, again, take shades of colour. It would be easy to find three stuffs of such closely similar shades that no difference could be perceived between the first and second, nor yet between the second and third, while yet the first and third would be distinguishable. In such a case, the second shade cannot be the same as the first, or it would be distinguishable from the third; nor the same as the third, or it would be distinguishable from the first. It must, therefore, though indistinguishable from both, be really intermediate between them. (*OKEW*, p. 142)[13]

What Russell sees as important in each of these cases is that our capacities for discrimination are outrun by changes in the items or qualities that we are experiencing such that we might erroneously conclude that there *are* no differences or boundaries between them. It is important to recognize that Russell is not here merely contending that we cannot distinguish between distinct qualities that are indistinguishably *qualitatively similar*. He is claiming that we cannot determine the *numerical identities* of the 'particularized qualities' (as we

[13] It is worth comparing these passages with those in *POP* involving the fading sound of a horse trotting away, transitions of one shade of colour to another, and the tuning of a musical instrument, which Russell uses to illustrate that the self-evidence of our intuitive judgements comes in degrees. For instance, he says, 'suppose we are comparing two shades of colour, one blue and one green. We can be quite sure they are different shades of colour; but if the green colour is gradually altered to be more and more like the blue, becoming first blue-green, then a greeny-blue, then blue, there will come a moment when we are doubtful whether we can see any difference, and then a moment when we know that we cannot see any difference' (pp. 215–6). While Russell here only notes that cases of this kind give rise to *doubt* as to whether two shades are different and does not make the stronger claim that similar shades can appear to be the same although they are different, it is clear that he would allow for this once we consider *why* these cases raise doubt: we cannot tell whether the shades are the same or different.

 © Wishon 2016

might call them) with which we are acquainted in sensation, nor of the particulars that have those qualities—in this case sense-data.[14]

After all, if the only thing at stake between Russell and the British Monistic Idealist here were indistinguishable qualitative similarity, his argument for ontological pluralism would appear to be largely, if not entirely, unsuccessful. And there is good reason to think that Russell appreciates this fact. On the other hand, if he is right that phenomenal continua cases bear on the question of whether we can determine the numerical identities of the objects of our acquaintance, he would thereby have firm grounds for rejecting the Thesis of Interpenetration as the only ontological view 'adequate to the senses'.

Charity is not the only evidence available for determining Russell's own view about the significance of phenomenal continua cases. On the contrary, Russell's own remarks demonstrate that he recognizes that *even in the case of our own sense-data*, we can fail to determine the numerical identities of the objects of our acquaintance, and thus acquaintance simply does not secure a subject against the possibility of misidentifying the objects of his or her thought and talk.

The problem with the Thesis of Interpenetration, as Russell diagnoses it, is that it ultimately rests on the supposition that 'the colours, being immediate data, must *appear* different if they *are* different' (*OKEW*, p. 144, emphasis in original). This, he urges, is a mistake, one which he thinks likely results from misunderstanding the nature of acquaintance and its epistemic features. In particular, he believes that the mistake is the result of conflating a subject's knowledge of a thing by *acquaintance* with the subject's possession of *knowledge of truths* about it. He remarks that:

> It is unconsciously assumed, as a premise for *reductio ad absurdum* of the analytic view, that, if A and B are immediate data, and A differs from B, then the fact that they differ must also be an immediate datum. It is difficult to say how this assumption arose, but I think it is to be connected with the confusion between 'acquaintance' and 'knowledge about'. Acquaintance, which is what we derive from sense, does not, theoretically at least, imply even the smallest 'knowledge about', *i.e.* it does not imply knowledge of *any* proposition concerning the object with which we are acquainted. (*OKEW*, p. 144, emphasis added)

[14] See page 212 of Russell (1911) for evidence that he distinguished between particularized qualities and the universals they exemplify. Also see page 324 of his 'On Meaning and Denotation' (1903) in which he explicitly states that there are 'particularized relations'. It is worth noting that this distinction appears to be in play, if only implicitly, in chapter ten of *POP*.

 © Wishon 2016

What's more, when such conflations of acquaintance and knowledge of truths are made, we are likely to be misled regarding acquaintance in a variety of ways. For instance, he says:

> It is a mistake to speak as if acquaintance had degrees: there is merely acquaintance and non-acquaintance. When we speak of becoming 'better acquainted', as for instance with a person, what we must mean is, becoming acquainted with more parts of a certain whole; but the acquaintance with each part is either complete or non-existent. Thus it is a mistake to say that if we were perfectly acquainted with an object we should know all about it. 'Knowledge about' is knowledge of propositions, which is not involved necessarily in acquaintance with the constituents of the propositions. (*OKEW*, pp. 144–5)

For Russell, the fact that we can have 'perfect' and 'complete' knowledge of a thing by acquaintance without knowing any truths about that thing is of great importance. One reason for this is that it would help to explain how commonsense and especially science can extend our knowledge of the world in a piecemeal fashion.[15] But another reason is that it allows him to directly challenge the Thesis of Interpenetration by denying that acquaintance secures us against making erroneous judgements about the numerical identity or diversity of its objects. On this latter point, Russell insists that:

> *To know that two shades of colour are different is knowledge about them; hence acquaintance with the two shades does not in any way necessitate the knowledge that they are different.* (*OKEW*, p. 145, emphasis added)

And a few pages later, he further drives home what he sees as the crucial significance of phenomenal continua cases:

> Is there, in actual empirical fact, any sufficient reason to believe the world of sense continuous? … if we see a coloured surface whose colour changes gradually, its sensible appearance if the change is continuous will be indistinguishable from what it would be if the change were by small finite jumps. If this is true, as it seems to be, it follows that there can never be any empirical evidence to demonstrate that the sensible world is continuous, *and not a collection of a very large finite number of elements* of which each differs from its neighbour in a finite though very small degree. (*OKEW*, p. 148, emphasis added)

[15] For more on this point, see Proops (2014) and Proops (2015). I would like to thank an anonymous referee who rightly points out that 'Russell is here pursuing his campaign (now largely forgotten by those who think about his epistemology) against neo-Hegelian monism. He is arguing for an atomist epistemology against Hegelian holists who hold that, abstracted from the Absolute, we have only a "partial and one-sided" knowledge of (say) a colour. In 1912, when he made the remark, these were still Russell's main enemies'.

 © Wishon 2016

Taken altogether, these passages provide powerful evidence that Russell in fact *explicitly rejects* the received view of acquaintance (or at least that he did reject this view at the time of his 1914 Lowell Lectures). They show that Russell thinks that we *can* and regularly *do* misidentify the objects of our acquaintance. More specifically, they show that a subject's sensory evidence does not always provide sufficient grounds for excluding either the possibility that sensory items *a* and *b* are identical or the possibility that they are distinct, both of which are equally live options. So we cannot know, simply on the basis of being acquainted with *a* and *b*, whether we are acquainted with two distinct things or one thing twice over. But more than that, Russell holds that even if we designate *a* and *b* with co-referring expressions the meanings of which we understand, and then consider whether *a* is *b*, we might not know what the right answer is. That is why logical analysis is on equal footing with British Monistic Idealism, empirically speaking.

At this point, a defender of the received view of acquaintance might argue as follows:[16] Perhaps it is true that Russell's discussions of phenomenal continua cases show that he rejects the assumption that where *a* and *b* are objects of sensory acquaintance which in fact differ, this difference must itself be immediately apparent. However, phenomenal continua cases do not quite contradict the central assumption of the received view that where *a* and *b* are objects of sensory acquaintance which are in fact identical, this identity must itself be immediately apparent, so that there is no serious question as to whether *a* is *b*. Indeed, a critic might continue, when it comes to judgements about numerical differences between objects it is understandable that knowledge of a colour need not be taken to include knowledge of all other colours from which it differs—but because true identities do not involve other objects, the same point does not apply.

Despite the initial plausibility of this line of argument, it does not hold up to scrutiny. For one thing, it does not square with Russell's remarks that 'the question whether sense-data are composed of mutually external units is not one which can be decided by empirical evidence' (*OKEW*, p. 143). If the received view of acquaintance were true, then in every case in which *a* is in fact *b* it would be immediately apparent that this is so. But then, conversely, whenever there is a serious empirical question of whether *a* is *b* (as there is in phenomenal

[16] I am grateful to Thomas Baldwin for advancing this line of argument against my reading of *OKEW*. Proops (2015) presses the same basic worry, as has Ralf Bader in conversation.

 © Wishon 2016

continua cases) we could straightforwardly conclude that *a* could not be *b*. Consequently, the issue of whether sense-data are composed of mutually external units *could* be settled by the empirical evidence, contrary to what Russell says. Hence charity suggests that we interpret him as allowing for the possibility of getting both false and true identities wrong.

For another thing, Russell's *stated* grounds for allowing for the possibility of errors in identification regarding the objects of acquaintance do not involve the issue that knowing numerical differences requires knowing more than one object. Rather, they are of a more general sort. As the passages from *OKEW* above indicate, Russell sees the assumption that if *a* and *b* are distinct the difference must be immediately apparent as connected with the conflation of knowledge by acquaintance with knowledge of truths, which he takes to be logically distinct species of knowledge (as we will see in §4 below). Knowledge of the numerical diversity of *a* and *b* is an instance of knowledge of truths about *a* and *b*, and is not implied by a subject's sensory acquaintance with *a* and *b*. Likewise, knowledge of the numerical identity of *a* and *b* is an instance of knowledge of truths—'*a* is *b*' is a proposition regardless of its truth or falsity. So if this logical distinction between knowledge by acquaintance and knowledge of truths is an important part of the explanation of how a subject can wrongly judge *a* and *b* to be identical when they are distinct, parity of reasoning suggests that there will be the possibility of misjudging *a* to be distinct from *b* when they are identical.

And thirdly, phenomenal continua cases can be easily fleshed out in such a way as to show that it is possible that a subject is acquainted with *a* and *b*, *a* is in fact *b*, and yet the subject misjudges that *a* is distinct from *b*. For instance, we can imagine a series of trials in which the subject is presented with pairs of colour patches which closely resemble each other.[17] In the first trial, the subject is presented with indiscriminably different colour patches *a* and *c* such that the subject wrongly judges that *a* is *c*. In the second trial, the subject is presented with indiscriminably different colour patches *b* and *d* such that the subject wrongly judges that *b* is *d*. In the third trial, the subject is presented with *c* and *d*, which are discriminably different from each other and the subject correctly judges them to be so. Now if the subject is queried about whether *a* is *b*, it is quite likely that the subject will draw on his or her previous identity judgements and conclude

[17] I am grateful to Bryan Harper for helpful discussion of this case.

 © Wishon 2016

that *a* is not *b* even if *a* is in fact *b*. Whatever the case may be, it is hard to deny that there will be an *open question* for the subject, one not settled by his or her sensory acquaintance with *a* and *b*, concerning whether *a* is *b*, even if *a* is in fact *b*. And the fact that there can so much as be an open question in phenomenal continua cases about the truth of a true identity statement concerning objects of acquaintance is enough to show that Russell does not plausibly intend acquaintance to safeguard us from cases of misidentification.[18]

4. Knowledge of things and knowledge of truths

Despite the clear evidence presented above, the received view of acquaintance is so familiar and established in philosophical lore that it might strike many philosophers as preposterous to suggest that Russell never held it.[19] Nevertheless, I believe that if we pay careful attention to what Russell actually *says* about acquaintance and its role in reference and cognition it should come as no surprise that he allows for the possibility of misidentifying the objects of acquaintance. So let us consider what Russell actually says about acquaintance.

On Russell's picture of reference and cognition, there is a fundamental distinction, one that is often marked in natural language, between two kinds of knowledge we can have about objects and their features. There is our familiar notion of *knowing that* such-and-such is the case. This sort of knowledge paradigmatically involves

[18] Proops (2011) too recognizes that nothing in Russell's views plausibly preclude errors of identification in cases in which a subject is twice presented with one and the same colour, feeling of weight, etc., over a short interval of time, particularly given his views about the duration of sense-data (which he allows to last for several seconds or even minutes) as well as the various kinds of acquaintance subjects can bear to things (including present acquaintance with past sense-data through memory). We can also imagine an analogous synchronic example with the help of Austin's (1990) case of viewing the same colour patch through two tubes. Given that in normal cases a subject can be visually acquainted with one and the same visual object twice-over with each eye, we can imagine cases in which the same subject is visually acquainted with the same patch twice over by viewing it with each eye through separate tubes while mistakenly thinking they are distinct patches due to the misconception that the two tubes are pointed in different directions. Austin himself does not think Russell would allow for such a case, but this is largely due to the misconception that Russell thinks sense-data are mental items. In point of fact, in both *POP* and *OKEW* Russell insists that sense-data are plausibly *external* to the mind—even if we cannot strictly prove that they exist when they are not sensed. Thus, it is hard to imagine a non-question-begging objection to such a case. See Wishon (2015) for an account of how Russell's views about the nature of acquaintance and its objects develop over the course of his body of work.

[19] Campbell, for instance, humorously exclaimed 'He's mad!' when I brought Russell's remarks to his attention.

 © Wishon 2016

conceptually articulated beliefs and is propositional (or multiply-relational) in character; it involves cognitive attitudes that can be evaluated in terms of their truth or falsity. When we have this kind of knowledge about an object, we know (or at least believe) that the object is thus-and-so. However, Russell points out that there is another way in which we use the term 'knowledge' in ordinary language: we talk of *knowing* objects in the sense that we know, whether directly on the basis of experience or indirectly by having a descriptive conception of it, what object it is. When we have this kind of knowledge about an object or feature, we simply know *it*.

Following William James, and others before him, Russell argues that any adequate philosophical theory of our knowledge must be careful to make an analogous distinction between our *knowledge of things* and our *knowledge of truths* about those things.[20] Indeed, in *POP*, Russell remarks:

> The word 'know' is … used in two different senses. (1) In its first use it is applicable to the sort of knowledge which is opposed to error, the sense in which what we know is *true*, the sense which applies to our beliefs and convictions, i.e. to what are called *judgements*. In this sense of the word we know *that* something is the case. This sort of knowledge may be described as knowledge of *truths*. (2) In the second use of the word 'know' above, the word applies to our knowledge of *things*, which we may call *acquaintance*. (p. 69, emphasis in original)

Knowledge of things, for Russell, is the basis from which all knowledge of truths about the world becomes possible. In its most basic form, our knowledge of things is the sort of knowledge that is constituted by our direct experiential contact, or acquaintance, with objects and features.[21] Knowledge of truths, on the other hand, is knowledge about (or at least belief about) facts or propositions involving such objects and features.

Russell conceives of acquaintance as a *relation* of direct awareness we bear to objects and features when they are presented to us in

[20] See chapter eight of James (1890/1950). To my knowledge, the earliest explicit discussion of this distinction appears in part one of John Grote's 1865 *Exploratio Philosophica: Rough notes on modern intellectual science*.

[21] Note that Russell is not entirely consistent in his uses of the terms 'acquaintance' and 'knowledge of things', even within the same work. Sometimes he identifies them, but other times he treats acquaintance as one kind of knowledge of things, in which case he allows for the possibility of knowledge of things by description. Given that knowledge by description puts one in a position to think and talk about things, but does not by itself assert that something is thus-and-so, I prefer the latter practice.

 © Wishon 2016

experience. Acquaintance is relational in the straightforward sense that it is a 'dual-relation' that obtains between a subject and the objects or features with which the subject is acquainted; both the subject and the objects or features are constituents of the acquaintance relation (or, more properly, of the relational fact). He is very clear in making this point on a number of occasions. For instance, in his 1913 manuscript *Theory of Knowledge* (hereafter *TOK*), Russell makes the following proposal:

> Now, since we have decided that experience is constituted by a relation, it will be better to employ a less neutral word; we shall employ synonymously the two words 'acquaintance' and 'awareness', generally the former. Thus when *A experiences* an object *O*, we shall say that *A* is acquainted with *O*. (Russell 1992a, p. 35, emphasis added)[22]

In addition, he makes the claim in his earlier 1911 paper entitled 'Knowledge by Acquaintance and Knowledge by Description' (hereafter KAKD) that:

> I am *acquainted* with an object when I have a direct cognitive relation to the object, i.e. when I am directly aware of the object itself. When I speak of a cognitive relation here, I do not mean the sort of relation which constitutes judgment, but the sort which constitutes presentation. (p. 148, emphasis in original)

Russell again characterizes acquaintance in roughly this way, though in more general terms, in chapter four of *POP*, where he says:

> The faculty of being acquainted with things other than itself is the main characteristic of a mind. Acquaintance with objects essentially consists in a relation between the mind and something other than the mind; it is this that constitutes the mind's power of knowing things. (pp. 66–7)

For Russell, it is extremely important that acquaintance is a *direct* conscious awareness, or 'presentation', of things. By calling acquaintance 'direct', he means to emphasize that such awareness of things, in contrast to our knowledge of things by description, is not inferentially or descriptively mediated. As he says in chapter five of *POP*:

> We shall say that we have *acquaintance* with anything of which we are directly aware, without the intermediary of any process of inference or any knowledge of truths. Thus in the presence of my table I am acquainted with the sense-data that make up the appearance of my table—its colour, shape, hardness, smoothness, etc.; all these are things of which I am

[22] For similar remarks, also see *TOK* (pp. 7–8, 35, 39).

 © Wishon 2016

> immediately conscious when I am seeing and touching my table. (p. 73, emphasis in original)

One important reason why Russell characterizes the directness of acquaintance in this way is that he sees it as the most epistemically *fundamental* cognitive relation to objects, one which constitutes our most basic form of intentionality, or object-directedness. He contends:

> All our knowledge, both knowledge of things and knowledge of truths, rests upon acquaintance as its foundation. (*POP*, p. 75)

In claiming this, Russell is not here endorsing any form of Cartesian epistemic 'foundationalism'.[23] There is compelling evidence that he rejects such views throughout his work. For instance, in chapter thirteen of *POP* he asserts that 'all our knowledge of truths is infected with *some* degree of doubt, and a theory which ignored this would be plainly wrong' (p. 210, emphasis in original). (It is worth noting that if knowledge of numerical identity is knowledge of truths, and *all* knowledge of truths is infected with uncertainty, then it straightforwardly follows that the received view of acquaintance is false.)

Instead, what Russell has in mind is that the knowledge we have of things when we are acquainted with them plays a fundamental role in our cognition of them. In particular, Russell's view is that our conscious awareness of things and their features is what puts us in a position to select and attend to them, to direct our information-gathering methods and capacities at them, and thereby to form a conception of them that, if all goes well, constitutes knowledge of truths about them. Thus, acquaintance is an enabling condition for our knowledge of truths about objects and conceptions of objects—which we can then exploit to think of objects indirectly by description.

Nevertheless, acquaintance, all by itself, is not a sufficient condition for our having *any* conception of, or knowledge of truths about, the things with which we are acquainted. In this way, our knowledge of things by acquaintance is markedly different from our indirect knowledge of things by description. Russell makes very clear the logical independence of our knowledge of things by acquaintance from any knowledge of truths about them:

> Knowledge of things, when it is of the kind we call knowledge by *acquaintance*, is essentially simpler than any knowledge of truths, and logically independent of knowledge of truths, though it would be rash to

[23] Russell's epistemic views are foundationalist only in the sense that our self-evident (but not infallible or absolutely certain) intuitive *knowledge of truths*, including logical truths, is the epistemic foundation for our derivative or inferred knowledge of truths (*POP*, pp. 170–3).

 © Wishon 2016

> assume that human beings ever, in fact, have acquaintance with things without at the same time knowing some truth about them. Knowledge of things by *description*, on the contrary, always involves … some knowledge of truths as its source and grounds. (*POP*, pp. 72–3, emphasis in original)

Consequently, we must be careful not to equate acquaintance with, or analyse it in terms of, our possession of any conception of the objects of our acquaintance, nor with any knowledge of truths about them.

Unfortunately, many have failed to recognize just *how* radically different knowledge of things and knowledge of truths are on Russell's view.[24] One reason for this is that many have been misled by passages such as the following oft-cited one from *POP*:

> The particular shade of colour that I am seeing may have many things said about it—I may say that it is brown, that it is rather dark, and so on. But such statements, though they make me know truths *about* the colour, do not make me know the colour itself any better than I did before: so far as concerns knowledge of the colour itself, as opposed to knowledge of truths about it, I know the colour perfectly and completely when I see it, and no further knowledge of it itself is even theoretically possible. Thus the sense-data which make up the appearance of my table are things with which I have acquaintance, things immediately known to me just as they are. (pp. 73–4, emphasis in original)

Russell's description here of knowing colours we are acquainted with 'perfectly and completely' and 'just as they are' has suggested to many that the crucial difference between knowledge of truths and knowledge of things is simply that the former is partial and fallible whereas the latter is in some important sense not.

However, there are compelling reasons for thinking that such readings of Russell misconstrue what he *means* by claiming that our knowledge of things does not allow for error. Russell's actual view becomes clearer later in chapter twelve of *POP* (among other places) when he remarks:

> Our knowledge of truths, unlike our knowledge of things, has an opposite, namely *error*. So far as things are concerned, we may know them or not know them, but there is no positive state of mind which can be described as erroneous knowledge of things, so long, at any rate, as we confine ourselves to knowledge by acquaintance. Whatever we are acquainted with must be something; we may draw wrong inferences from our acquaintance, but acquaintance itself cannot be deceptive. Thus there is no dualism as regards acquaintance. But as regards knowledge of truths, there is a

[24] For some exceptions, see Wahl (2007), Campbell (2009), Proops (2011), and Amijee (2013).

 © Wishon 2016

> dualism. We may believe what is false as well as what is true. (p. 186, emphasis in original)

On Russell's view, there simply is no sense to be made of acquaintance being erroneous because it simply is not the sort of epistemic relation that provides a 'take' on things. For unlike our knowledge of truths, our knowledge by acquaintance does not in any way purport that things are thus-and-so and so cannot purport that things are otherwise than they are. Rather, acquaintance is simply the direct experiential contact with things that puts us in a position to take them as being thus-and-so.

We can usefully make Russell's point as follows: in the case of our knowledge of truths, we can not only have or not have the relevant cognitive states; when we do have them they can be true or false. But in the case of acquaintance, we can only have it or not have it; there simply is no further issue of truth or falsity. And this is precisely what Russell really seems to be getting at when, in a passage cited in the previous section of this paper, he says: 'it is a mistake to speak as if acquaintance had degrees: there is merely acquaintance and non-acquaintance' and 'acquaintance with each part [of a thing] is either complete or non-existent' (*OKEW*, pp. 144–5).

Still, one might grant that acquaintance is logically distinct from any knowledge of truths while nevertheless insisting that having acquaintance with an object puts a subject in a position sufficient to rule out their making erroneous judgements about at least some facts about the object, *were* the subject to make them. So *if* Russell thinks that judgements about the identities of the objects of acquaintance are ones whose infallibility is ensured by, even if not constituted by, a subject's acquaintance, then the received view would be vindicated.

To be sure, Russell makes several remarks in *POP* which lend a good deal of credibility to such an interpretation. For instance, at the end of chapter four he says, 'If I am acquainted with a thing that exists, my acquaintance gives me knowledge that it exists' (*POP*, p. 70). In closing chapter ten, he claims that 'whatever may be the object of acquaintance, even in dreams and hallucinations, there is no error involved so long as we do not go beyond the immediate object: error can only arise when we regard the immediate object, i.e. the sense-datum, as the mark of some physical object' (*POP*, p. 172). In chapter eleven, Russell proposes that sensory acquaintance with an object puts a subject in a position to form an intuitive judgement of perception that it exists, one which has the highest degree of

 © Wishon 2016

self-evidence (*POP*, pp. 179–84). He further goes on to say that there is a sense of 'self-evidence' corresponding to this highest degree which 'is really an infallible degree of truth' (*POP*, pp. 184–5). In chapter thirteen, Russell similarly states:

> In all cases where we know by acquaintance a complex fact of certain terms in a certain relation, we say that the truth that these terms are so related has the first or absolute kind of self-evidence, and in these cases the judgment that the terms are so related must be true. Thus this sort of self-evidence is an absolute guarantee of truth'. (*POP*, pp. 213–4)

Finally, in chapter fourteen he insists that 'some knowledge, such as the knowledge of the existence of our sense-data, appears quite indubitable, however calmly and thoroughly we reflect upon it' (*POP*, p. 235). Together with Russell's claim that a subject knows an object 'perfectly and completely' when he or she is acquainted with it, such passages convince many that Russell did indeed embrace the received view.

Nevertheless, there are compelling reasons for resisting this conclusion. For one thing, even if Russell *were* to think that sensory acquaintance positions a subject to make infallible judgements about the *existence* of the objects of acquaintance, more argument would be needed to show that such knowledge includes infallible knowledge of truths about their *identities*. Yet Russell's discussions of phenomenal continua cases provide powerful grounds for thinking that he would *not* have included the latter kind of knowledge as infallible regardless of whether the former is.

For another thing, Russell makes a number of remarks elsewhere in *POP* which stand in direct tension with standard readings of the passages regarding intuitive judgements, absolute self-evidence, and infallibility. For instance, we have already seen that Russell insists that 'all our knowledge of truths is infected with *some* degree of doubt, and a theory which ignored this fact would be plainly wrong' (*POP*, p. 210, emphasis in original). Similarly, at the end of chapter two he contends 'it is of course *possible* that all or any of our beliefs may be mistaken, and therefore all ought to be held with at least some slight element of doubt' (*POP*, pp. 39-40, emphasis in original). And as Proops (2015) has recently noted, immediately after Russell claims that the kind of self-evidence which accompanies a subject's acquaintance with a complex fact is an absolute guarantee of truth, he clarifies that 'although this sort of self-evidence is an absolute guarantee of truth, it does not enable us to be *absolutely* certain, in the case of any given judgement,

 © Wishon 2016

that the judgement in question is true' (*POP*, p. 214). Hence, it is very unclear that Russell takes *any* of our judgements to have the kind of infallibility required for a defence of the received view.

While it is beyond the scope of this paper to settle these complex issues, I am inclined to agree with Proops that on Russell's somewhat carelessly stated view, 'to say that *S*'s perceiving a complex is an absolute guarantee of the truth of the corresponding judgement is not to say (and neither does it entail) that when *S* perceives the complex in question *S* thereby knows with certainty that the corresponding judgement is true' since *S* might well erroneously form a judgement that does not correspond to the fact with which he or she is acquainted (2015, p. 52). Moreover, I am also inclined to agree with Proops that rather than taking Russell to hold that acquaintance with an object positions us to have infallible knowledge of truths about its existence, he is better understood as saying that such acquaintance is accompanied by *knowledge by acquaintance* with the fact of the object's existence—and hence there is no issue of truth or error (2015, p. 6). Of course, if the subject were to form the judgement that the object exists in such circumstances, the judgement could not fail to be true given that the subject could not be acquainted with the existence fact and it not be the case. I grant, however, that there is a great deal more that could be said on these controversial matters.[25]

5. Knowledge by acquaintance and proper names

Setting aside these issues about the relationship between knowledge by acquaintance and knowledge of truths, one might still wonder how Russell *could* reject the received view of acquaintance wholesale given his views about the nature of judgements at various points in his writings.[26] After all, it is widely recognized that on his early views on judgement he takes the proposition expressed by '$a = b$', if true, to be one and the same proposition as the one expressed by '$a = a$'. And on his views from 1912-1917, he takes a subject making the

[25] See Amijee (2013), Linsky (2015), and Wahl (2015) for further considerations in favour of this reading.

[26] I am grateful to an anonymous referee for raising this worry. It is worth pointing out that the worry here is even more pressing for Russell after he gives up on in 1919 propositions, the multiple-relation theory of judgement, *and* the act-object distinction, since if there is no distinction between the subject making the judgement and the objects of acquaintance, there seems to be even less room for errors of identification. Nevertheless, there is strong evidence that Russell thinks errors of identification are still possible, especially in 'Vagueness' (1923).

 © Wishon 2016

judgement '$a = b$', if true, to be multiply-related to the very same elements as he or she would be in making the judgement '$a = a$'. So if a subject's judgement includes as constituents only those elements that he or she is acquainted with, it certainly appears as if Russell could not countenance differences in the meanings of '$a = a$' and '$a = b$', nor allow that a subject could fail to know whether '$a = b$' on consideration if he or she understood the meaning of 'a', 'b', and '='.

As compelling as this worry appears at first glance, it only works given certain assumptions about the nature and epistemology of judgements, assumptions that Russell himself seemingly does not accept. In particular, I will argue, this worry depends on the assumption that if a subject makes a judgement and understands its meaning, the judgement's meaning (and logical form) will be *fully transparent* to the subject, presumably because the subject has discriminating conceptions of its constituents. However, there is good reason to think that Russell never accepted such a strong link between epistemic matters and the meaning of a judgement or expression. To see this, we must carefully examine Russell's views about the relationship between knowledge by acquaintance and the meanings of referring expressions. What we will find is that Russell ties referring expressions to acquaintance for purposes quite different from trying to solve Frege's Puzzle in full generality.

In point of fact, Russell's primary aim is to explain the fundamental role that conscious experience plays in reference and cognition. In particular, Russell wants to explain how our direct conscious awareness of things puts us in a position to make *genuine reference* to them in thought and talk, as well as how we can derivatively come to think and talk significantly about things with which we lack direct experiential contact.

Russell holds the view that our thoughts, and the statements that express them, are associated with conditions such that the thoughts and statements are true if the conditions obtain, and false if they do not. Furthermore, he thinks that the objects and features that genuinely referring devices pick out are part of these conditions. With 'denoting' devices, in contrast, the objects or features denoted are picked out by these conditions, and required to exist if the conditions obtain. But they are not themselves integral parts of these truth-conditions. Hence, the distinctive feature of a genuinely referring expression is that its *meaning* (that is, its contribution to the truth-conditions of thoughts or utterances in which it occurs) is the very worldly object or feature, if any, for which it stands. As Russell puts it in KAKD:

 © Wishon 2016

> The denotation [of an expression], I believe, is not a constituent of the proposition, except in the case of proper names, i.e. of words which do not assign a property to an object, but merely and solely name it. (p. 157)

What's more, unlike in cases of 'denoting' expressions, such as definite descriptions, in the case of genuinely referring expressions, we need not have any conception of their referents, nor of the mechanisms whereby their reference is determined.

We might draw the distinction here as one between what we might call '*reference-guided conceptions of things*' and '*conception-guided reference to things*'. In 'The Philosophy of Logical Atomism' (hereafter PLA), Russell makes quite explicit that he views the former as more fundamental:

> When you have acquaintance with a particular, you understand that particular itself quite fully, independently of the fact that there are a great many propositions about it that you do not know, *but propositions concerning the particular are not necessary to be known in order that you may know what the particular itself is.* It is rather the other way round. In order to understand a proposition in which the name of a particular occurs, you must already be acquainted with that particular. The acquaintance with the simpler is presupposed in the understanding of the more complex. (pp. 65–6, emphasis added)

In this way, Russell holds that *all* cognition and linguistic designation ultimately rests more or less directly on our acquaintance with, or conscious awareness of, objects and their features. This is because acquaintance not only puts us in a position to make genuine reference to things (and understand the meanings of such expressions), it also puts us in a position to pick up information about them and thereby form conceptions of them. In turn, these conceptions are precisely what enable us to describe, and thereby think and talk meaningfully about, things with which we lack experiential contact. We accomplish this in the following way:

> What happens, in cases where I have true judgement without acquaintance, is that the thing is known to me by *description*, and that, in virtue of some general principle, the existence of a thing answering to this description can be inferred from the existence of something with which I am acquainted. (*POP*, p. 71, emphasis in original)

Simply put, our knowledge of things by acquaintance is an essential part of the explanation of *how it is* that we acquire the conceptual materials, both about particular things and the general way the world works, required for conception-guided designation in the first place.

 © Wishon 2016

Russell's acceptance of 'Russell's Principle', *weakly* construed, is fully explained by these sorts of considerations, a point made more clearly in his statement of it in KAKD:

> *Every proposition which we can understand must be composed wholly of constituents with which we are acquainted* … The chief reason for supposing the principle true is that it seems scarcely possible to believe that we can make a judgment or entertain a supposition without knowing what it is that we are judging or supposing about. (p. 154, emphasis in original)

And again later when he stresses that:

> All propositions intelligible to us, whether or not they primarily concern things only known to us by description, are composed wholly of constituents with which we are acquainted, *for a constituent with which we are not acquainted is unintelligible to us.* (KAKD, p. 161, emphasis added)

These remarks suggest that, for Russell, the crucial issue is not whether acquaintance puts a subject in a position to distinguish the referents of genuinely referring expressions *from all other things.* Rather it is simply that understanding the meanings of genuinely referring expressions, and derivatively of descriptive phrases composed of them, requires knowing what objects they have been conventionally assigned to designate. And we must be careful not to conflate a subject's *understanding* a proposition involving genuinely referring expressions with the subject's being precluded from making errors when identifying the objects designated, for as Russell notes, 'we can often understand a proposition in cases where we have not enough knowledge to make a judgment' (KAKD, p. 155).

It is worth noting that this is an issue that Russell returns to on many occasions, perhaps most vividly in his 1957 'Mr. Strawson on Referring' (hereafter MSOR).[27] In response to P.F. Strawson's challenge that there are no genuinely referring expressions, in Russell's

[27] A brief aside: it is often maintained that the later Russell is an unreliable authority about his earlier views and so is not to be trusted on that account. I believe that his alleged unreliability has been largely overstated. There is no denying that Russell was mistaken on various occasions about specific details having to do with when and why he made certain changes to his earlier views. But it does not follow that he was mistaken about broad features of his view, especially given that he states them on many occasions throughout his career, even if in slightly different terms. Moreover, it is one thing to take Russell as unreliable about specific *changes* in his views, but it is quite another to presume, on that basis, and quite uncharitably, that he is unreliable regarding aspects of his views he claims *never* to have changed, especially when there are available interpretations which would corroborate his reports.

 © Wishon 2016

sense, in ordinary language (nor any definite descriptions), Russell replies:

> So far as I can discover from the context, what [Strawson] objects to is the belief that there are words which are only significant because there is something they mean, and if there were not this something, they would be empty noises, not words … The necessity of such words is made obvious by ostensive definition … Unless fundamental words in the individual's vocabulary had this kind of direct relation to fact, language in general would have no such relation. (MSOR, p. 632)

Once again, the key issue here is not Frege's Puzzle or substitution problems; it is the issue of how language, at the fundamental level, relates to things and features in the world as well as how we come to understand expressions in the relevant language.

Still, one might reasonably wonder whether Russell is *consistent* in allowing, on the one hand, that a subject can misidentify objects of acquaintance when using expressions he or she understands, and on the other, that the only constituents involved in a true identity judgement of this sort are the subject, the relation of judging, the relation of identity, and the same object twice over. I believe that he *is* consistent on these matters. Russell, unlike many of his later interpreters, draws a clear distinction between what expressions mean, *logically speaking*, when they are used for logical purposes, and what ordinary expressions mean, *cognitively speaking*, when they are used in daily life. Without question, he takes the former to capture essential aspects of the latter. But he never pretends that it can give the whole story.

One place where we can clearly see this is again in Russell's reply to Strawson's charges that his Theory of Descriptions (and genuine proper names) 'embodies some fundamental mistakes' (Strawson 1950, p. 321) about the meanings of expressions in ordinary thought and talk, and that 'neither Aristotelian nor Russellian rules give the exact logic of any expression of ordinary language; for ordinary language has no exact logic' (Strawson 1950, p. 344). In response, Russell wholeheartedly agrees that 'ordinary language has no exact logic' yet maintains that the 'theory of descriptions was never intended as an analysis of the state of mind of those who utter sentences containing descriptions' (MSOR, p. 633). Quite the opposite, in OD he intends the Theory of Descriptions to solve certain *technical problems of logic*. And he insists that 'for technical purposes, technical languages differing from those of daily life are indispensable' (MSOR, p. 633). Indeed, he continues, 'common speech is full of vagueness and inaccuracy, and

 © Wishon 2016

... any attempt to be precise and accurate requires modification of common speech both as regards vocabulary and as regards syntax' (MSOR, p. 632). So, unless we uncharitably take Russell to be misreporting his earlier views and motivations, it is hard to deny that he draws the distinction between what expressions mean in logical analysis and what they mean in daily life around the time he writes OD.

Another important source of evidence that Russell distinguishes between the ordinary significance of our thought and talk and its logical significance can be found in his 1923 'Vagueness'. While the primary aim of the paper is to 'prove that all language is vague' (Russell 1923, p. 147) and so might appear not to address squarely the issue of whether acquaintance secures a subject against making errors of identification, Russell's discussions therein of phenomenal continua cases and the ultimate source of vagueness clearly show otherwise. Regarding the former, he asks us to:

> Consider the various ways in which common words are vague, and let us begin with such a word as 'red'. It is perfectly obvious, *since colours form a continuum*, that there are shades of colour concerning which we shall be in doubt whether to call them red or not, not because we are ignorant of the meaning of the word 'red', but because it is a word the extent of whose application is essentially doubtful. (Russell 1923, p. 148, emphasis added)

Clearly, one important lesson Russell wants us to draw from this is that all expressions in ordinary language standing for sensible qualities (including all expressions used in natural science for the quantitative measurement of observable phenomena) are vague. But what is more interesting is his explanation for why these expressions are vague: due to the limits of our physiological and cognitive constitution 'all observations have a margin of error' (Russell 1923, p. 149). In other words, Russell takes the vagueness of much of our thought and talk about sensible phenomena to be explained precisely by the physiologically based, coarse grained character of our discriminatory capacities for sensory stimuli (pp. 149–50). Hence the looseness of the meaning of our ordinary thought and talk is *explained by* the fact that sensory experience does not provide us with discriminating conceptions of its objects.

It is important to note that Russell takes this to apply to proper names and not simply predicates such as 'red'. In point of fact, he insists, 'the vagueness of the knowledge derived from the senses infects all words in the definition of which there is a sensible element' (Russell 1923, p. 150). Even more generally, Russell argues:

 © Wishon 2016

> The fact is that *all* words are attributable without doubt over a certain area, but become questionable within a penumbra, outside which they are again certainly not attributable. Someone might seek to obtain precision in the use of words by saying that no word is to be applied in the penumbra, but unfortunately the penumbra itself is not accurately describable, and all the vaguenesses which apply to the primary use of words apply also when we try to fix a limit to their indubitable applicability. (Russell 1923, p. 149, emphasis in original)[28]

Once again, it must be emphasized that the vagueness in our thought and talk is seemingly not simply the result of loose concepts which do not carve up the world as finely as our epistemic capacities do; rather, it is because our most basic epistemic contact with the world does not provide us with uniquely discriminating knowledge of it.

Lastly, it must be noted that in 'Vagueness' Russell explicitly diagnoses the Thesis of Interpenetration as the result of what he calls 'the fallacy of verbalism' (p. 147). This is very clear from his early remarks:

> There is a certain tendency in those who have realised that words are vague to infer that things are also vague. We hear a great deal about the flux and the continuum and the unanalysability of the Universe, and it is often suggested that as our language becomes more precise, it becomes less adapted to represent the primitive chaos out of which man is supposed to have evolved the cosmos. This seems to me precisely a case of the fallacy of verbalism—the fallacy that consists in mistaking the properties of words for the properties of things. Vagueness and precision alike are characteristics which can only belong to a representation, of which language is an example. They have to do with the relation between a representation and that which it represents. (Russell 1923, pp. 147–8)

Taking everything together, Russell's remarks about phenomenal continua cases, vagueness, and the distinction between meaning in the context of logic and in the context of daily life quite convincingly show that his views about the nature of judgements are perfectly consistent with his rejection of the received view of acquaintance, in all its versions.[29]

[28] It should be kept in mind that Russell is not using phrases such as 'without doubt', 'certainly', and 'indubitability' with strict literality, as the context of the discussion makes clear.

[29] For another statement of Russell's view that vagueness permeates all our thought and talk, see pages 37–8 of PLA in which he remarks 'I think vagueness is very much more important in the theory of knowledge than you would judge it to be from the writings of most people. Everything is vague to a degree you do not realize till you have tried to make it precise, and everything precise is so remote from everything that we normally think, that *you cannot for a moment suppose that is what we really mean when we say what we think*' (emphasis

 © Wishon 2016

6. Reconsidering 'On Denoting' and the puzzle of George IV

Before concluding, I must address the most serious challenge to my reading of Russell on acquaintance: it seems to stand in serious conflict with his treatment of the puzzle of George IV in OD, and elsewhere. Simply put, the puzzle of George IV appears to be a special instance of Frege's Puzzle in which the substitution of one genuinely referring expression with another seemingly co-referring expression licenses deductive inferences which intuitively strike us as invalid. The problem, as Proops (2011) has recently pointed out, is that if both expressions are indeed co-designating referring expressions, and both simply contribute their referents (and nothing else) to every proposition (or multiple relation) expressed by statements in which they occur, then evidently substitutions of one expression for the other should be entirely permissible *salva veritate*, at least given Russell's other commitments (pp. 155–9). Russell's Theory of Descriptions (and Descriptive Theory of Ordinary Names) is intended to explain why the relevant problematic substitutions are not really instances of substituting one propositional (or multiple-relational) constituent for itself, and in so doing to show that the appearance of a counterexample is merely illusory.

In OD, Russell presents the puzzle of George IV as follows:

> If *a* is identical with *b*, whatever is true of the one is true of the other, and either may be substituted for the other in any proposition without altering the truth or falsehood of that proposition. Now George IV wished to know whether Scott was the author of *Waverley*; and in fact Scott *was* the author of *Waverley*. Hence we may substitute *Scott* for *the author of 'Waverley'*, and thereby prove that George IV wished to know whether Scott was Scott. Yet an interest in the law of identity can hardly be attributed to the first gentleman of Europe. (p. 420)

The puzzle which Russell seems to be interested in here is this: If 'Scott' and 'the author of *Waverley*' both simply contribute the particular *Scott* to the truth-conditions of thoughts or utterances in which they occur, then given the nature of numerical identity, we should be able to substitute one for the other in any thought or utterance without altering its truth-value. But substituting 'Scott' for 'the author of *Waverley*' in 'George IV wished to know whether Scott was the author

---

added). See also his remarks in chapter ten of *AMi* that 'the meaning of a word is not absolutely definite: there is always a greater or less degree of vagueness' (pp.197–8), as well as his discussion of proper names in *HK* (pp. 87–9, 96–7).

 © Wishon 2016

of *Waverley*' does evidently change its truth-value, even though Scott did solely author *Waverley*.

Put slightly differently, the problem is this: if 'Scott' and 'the author of *Waverley*' are co-designating referring expressions, then the following argument (alluded to above) would be valid (Proops 2011, p. 155):

(1) Scott is the author of Waverley
(2) George IV wished to know whether Scott was the author of Waverley
(C) George IV wished to know whether Scott was Scott

As Proops notes, this argument involves 'a patently invalid substitutional inference', a point that is not lost on Russell (2011, p. 156). Hence, Russell aims to explain why this argument is invalid, and he sees the Theory of Descriptions as providing just the needed materials.

Russell's answer to the puzzle of George IV is simply to deny that 'the author of *Waverley*' *means* Scott, or anybody else (Russell 1959, p. 85). In other words, the descriptive phrase 'the author of *Waverley*' does not contribute *any* object or feature into the truth-conditions of thoughts or utterances in which it occurs. But it does not contribute a denoting concept, as he previously held, or a Fregean sense to them either. Instead, descriptive phrases such as 'the author of *Waverley*' are incomplete symbols, which contribute (restricted) existential quantifiers to the truth-conditions of thoughts or utterances in which they occur, but not necessarily the same ones on every occasion. One and the same descriptive phrase might, for instance, contribute a wide-scope (restricted) existential quantifier to the truth-conditions of one kind of thought or utterance, but a narrow-scope one to those of another (OD, pp. 424–5).

But while descriptive phrases do not uniformly contribute the *same* constituents to the truth-conditions of every thought or utterance in which they occur, it does not follow that they lack meaning altogether. On the contrary, descriptive phrases have meanings-in-context; we can give them a 'contextual definition' which specifies what they contribute to the truth-conditions of particular thoughts or utterances in particular contexts. In doing so, Russell maintains, the descriptive phrase will be 'broken up' in the sense that 'in any proposition in which it occurs the proposition, fully expressed, does not contain the phrase' (OD, p. 423). But so long as we can explain what a descriptive phrase contributes to the truth-conditions of a thought or utterance in the particular context in which it occurs, we can fully account for its

 © Wishon 2016

meaningfulness without taking it to stand, all by itself, for anything at all.

The upshot of the Theory of Descriptions for the puzzle of George IV is straightforward. Since 'Scott' and 'the author of Waverley' contribute different elements to the truth-conditions of assertions such as 'George IV wished to know whether Scott was the author of Waverley', it is no surprise that substituting one for the other does not preserve truth-value in all contexts. Nor is it surprising that substituting one for the other in the course of an argument can render it invalid. Thus, the Theory of Descriptions solves both aspects of the puzzle, without appealing to problematic denoting concepts or Fregean senses.

On the prevailing reading of OD, Russell intends this treatment of the puzzle of George IV—together with the Descriptive Theory of Ordinary names—to *generalize* to a wide range of instances of Frege's Puzzle (and other substitution problems). In other words, *whenever* we have an apparent instance of Frege's Puzzle, Russell's alleged strategy is to treat at least one of the expressions as a (often 'truncated') descriptive phrase which contributes, in the relevant context, a (restricted) existential quantifier to the truth-conditions of the relevant thought or utterance. On the flipside, he allegedly takes it that whenever we truly have a case of an identity sign flanked by co-designating referring expressions, instances of Frege's Puzzle (and other substitution problems) cannot arise. This is precisely because Russell embraces the received view of acquaintance. As Proops puts it:

> In order to avoid this difficulty [that is, George IV-style puzzles involving genuinely referring expressions], Russell must make substantive assumptions about either the nature of the referents of genuine Russellian names … or the nature of our access to them. That is to say, he must assume either (a) that Russellian Objects, by their very nature, lack 'facets' or 'sides', or (b) that, even if they have sides that would be hidden from a less perspicacious species of cognizers, our acquaintance is a searchlight sufficiently penetrating to lay those sides open to view. Either of these assumptions would guarantee the truth of the principle of Full Disclosure … and so would seem to preclude informative identities of the kind just described. (2011, pp. 166–7)[30]

[30] It is worth noting that some of Russell's own remarks about 'fleeting' sense-data might lead one to the relatively common misconception that he is committed to one-sided 'Russellian objects'. For instance, in PLA he says that '[sense-data] last for a finite time, a matter of seconds or minutes or whatever it may happen to be' (p. 65). However, it is doubtful that his remarks about fleeting sense-data are motivated by a desire to rule out

 © Wishon 2016

There certainly is a good deal of textual evidence in OD which lends credibility to this reading. For one thing, Russell first presents the puzzle of George IV in OD immediately after his criticisms of Frege's sense–reference distinction. So it is natural to see his solution to the former puzzle as a complete alternative to Frege's own answer to problems of substitution. For another, Russell explicitly states that 'the usefulness of *identity* is explained by [the Theory of Descriptions]. No one outside a logic-book ever wishes to say "x is x", and yet assertions of identity are often made in such forms as "Scott was the author of *Waverley*"' (OD, p. 426, emphasis in original). At face value, he certainly *seems* to be claiming here that the one and only non-trivial use of identity consists in its application to cases like George IV, where at least one of the expressions flanking the identity sign is a descriptive phrase ('truncated' or otherwise). In turn, this suggests that Russell sees all assertions of identity involving co-designating referring expressions as trivial, further vindicating the received view of acquaintance.

However, the most compelling evidence for the received view of acquaintance is provided, in my estimation, by what Stephen Neale helpfully calls the 'Triviality Argument', which appears in various forms in KAKD, PLA, *Principia Mathematica* (hereafter *PM*), and elsewhere (Neale 2005, pp. 824–5). In *MPD*, Russell gives perhaps his most succinct statement of this argument:

> The central point of the theory of descriptions was that a phrase may contribute to the meaning of a sentence without having any meaning at all in isolation. Of this, in the case of descriptions, there is precise proof: If 'the author of *Waverley*' meant anything other than 'Scott', 'Scott is the author of *Waverley*' would be false, which it is not. If 'the author of *Waverley*' meant 'Scott', 'Scott is the author of *Waverley*' would be a tautology, which it is not. Therefore, 'the author of *Waverley*' means neither 'Scott' nor anything else—i.e. 'the author of *Waverley*' means nothing at all, Q.E.D. (p. 85)

For our purposes, the most important premise in this argument is the claim that whenever we have an assertion with the genuine logical form 'a = b' it is either false or, if true, a trivial tautology. For Russell is clearly committing himself here to the claim that '"$a = b$"

---

problems of misidentification. Indeed, diachronic phenomenal continua cases can occur during very short periods of time, and synchronic cases are entirely unaffected. I am grateful to Russell Wahl for drawing my attention to Russell's remarks about fleeting sense-data.

 © Wishon 2016

is as trivial as "$a = a$" when a and b are both genuine names', as Neale puts it (2005, p. 825).

Further unequivocal evidence that Russell accepts the Triviality Argument is provided by the following remarks from PLA, which are worth quoting in full:

> I should like to make clear what I was saying just now, that if you substitute another name in place of 'Scott' which is also a name of the same individual, say, 'Scott is Sir Walter', then 'Scott' and 'Sir Walter' are being used as names and not descriptions, your proposition is strictly a tautology. If one asserts 'Scott is Sir Walter', the way one would mean it would be that one was using the names as descriptions. One would mean that the person called 'Scott' is the person called 'Sir Walter', and 'the person called "Scott"' is a description, and so is 'the person called "Sir Walter"'. So that would not be a tautology … But if you are using both as names, the matter is quite different. You must observe that the name does not occur in that which you assert when you use the name … So if I say 'Scott is Sir Walter', using these two names as names, neither 'Scott' nor 'Sir Walter' occurs in what I am asserting, but only the person who has these names, and thus what I am asserting is a pure tautology. (p. 114)[31]

Russell makes the same basic points regarding 'Scott' and 'Sir Walter' in chapter sixteen of his 1919 *Introduction to Mathematical Philosophy* (hereafter *IMP*). Hence, it is undeniable that Russell takes true assertions with the genuine logical form of '$a = b$' to be tautologies (at least he took them to be so from 1910 onwards). And if one assumes that the meaning of a trivial, tautological assertion of identity would be fully transparent to the subject making it, then by all appearances, the received view of acquaintance is on extremely solid ground. Nevertheless, these appearances are deceptive. Russell seemingly never accepted the received view, and nothing he says regarding the puzzle of George IV should lead us to believe otherwise.

The first thing that we must keep firmly in mind when considering how the puzzle of George IV might bear on our understanding of acquaintance is the point made in §4 above: Russell consistently distinguishes between the meanings of expressions when analysed for logical purposes and their meanings in ordinary thought and talk. It simply does not follow from his claim that true assertions of the genuine logical form '$a = b$' are trivial or tautological, logically speaking, that he takes them to be transparently true and cognitively trivial for the subject making them. In point of fact, we have already seen that Russell

[31] Mark Crimmins was the first to press me with these passages, though Peter Hylton, Kevin Klement, and an anonymous referee pointed them out as well. My thanks go to each of them.

 © Wishon 2016

denies outright that his analysis of logical meaning in OD is intended to capture faithfully the cognitive meaning of ordinary thought and talk.

But even setting this general worry aside, there are a number of more specific reasons for thinking that the puzzle of George IV in no way vindicates the received view of acquaintance. To begin with, even if Russell were to think that the Theory of Descriptions explains *all* instances in which 'identity is worth affirming', there is little reason to think that Russell took *acquaintance* to be part of the explanation for this (OD, p. 492). Indeed, it is striking that he never explicitly *says* that acquaintance shields a subject from the possibility of misidentifying the objects of acquaintance, either in OD or in subsequent discussions. Nor does he ever even claim explicitly that acquaintance plays a significant *role* in solving Frege's Puzzle, apart from its aforementioned role in all thought and talk.[32] In fact, he never so much as *mentions*

[32] For a few examples of Russell's subsequent discussions of the puzzle of George IV, see pages 158–61 of KAKD, part six of PLA, pages 173–80 of *IMP*, and pages 84–5 of *MPD*. One might object, as an anonymous referee has, that while Russell never *explicitly* touts his Theory of Descriptions as even a partial solution to Frege's Puzzle, he nevertheless plausibly has it squarely in mind as a target in KAKD. After all, immediately after he concludes that the descriptive phrase 'the man whose name was *Julius Caesar*' in the judgement expressed by 'the man whose name was *Julius Caesar* was assassinated' must be broken up 'if we are to be acquainted with all the constituents of the judgment', he says that 'this conclusion, which we have reached from considerations concerned with the theory of knowledge, is also forced upon us by logical considerations', at which point he proceeds to discuss the alleged distinction between '*meaning* and *denotation*'. But then, one might reason, is Russell not at least *suggesting* that his views about restricting the constituents of thought to the objects of acquaintance are motivated at least partly by Frege's Puzzle? And does Russell not at least *then* restrict acquaintance to sense-data, abstracta, and (perhaps) the self? So why not simply conclude that Russell holds the received view in KAKD?

While this line of reasoning might seem initially plausible, it does not hold up to scrutiny. For one thing, the conclusion that Russell is claiming to have reached from considerations of the theory of knowledge is that the phrase 'the man whose name is *Julius Caesar*' must be broken up if we are to be acquainted with all the constituents of judgments expressed by utterances containing it. And the basis for this conclusion is that we are not presented in experience with Julius Caesar, and so we cannot be acquainted with the person denoted by 'the man whose name is *Julius Caesar*'. But nothing here supports the received view of acquaintance.

Moreover, when Russell turns his attention to the logical considerations for breaking up the descriptive phrase and replacing it with a (restricted) existential quantifier, what he really wants to show is that identity statements involving descriptive phrases are assigned the wrong truth-conditions regardless of whether we take them to contribute their meanings *or* their denotations. Indeed, he argues, the *meanings* (or senses) of 'the author of Waverley' and 'the author of Marmion' are clearly not identical, so they cannot be what is contributed to the proposition expressed by 'the author of Waverley is the author of Marmion'. But, he contends, the *denotations* of these two descriptive phrases cannot be what is contributed either. For one thing, he says, we can understand 'the author of Waverley is the author of Marmion' without being acquainted with the denotation of either one, i.e. Scott. For another, it raises problems

 © Wishon 2016

acquaintance in the passages in OD in which he discusses the puzzle of George IV; nor does he mention the puzzle of George IV (or any misidentification or substitution problems) in the passages in which he explicitly discusses acquaintance (pp. 479–80, 492–3).

Moreover, there is strong evidence that at the time of *OD* Russell does not accept important aspects of his later epistemological views which are required for the received view of acquaintance to have any plausibility. In particular, he seemingly does not accept his later views that the objects of our acquaintance are special 'Russellian Objects' such as sense-data.

For instance, in his well-known 1904 'Letter to Frege' (which he cites without reservations or amendments in OD), he maintains:

> I believe that in spite of all its snowfields Mont Blanc itself is a component part of what is actually asserted in 'Mont Blanc is more than 4,000 metres high'. We do not assert the thought, for this is a private psychological matter: we assert the object of the thought, and this is, to my mind, a certain complex (an objective proposition, one might say) in which Mont Blanc is itself a component part. If we do not admit this, then we get the conclusion that we know nothing at all about Mont Blanc … In the case of a simple proper name like 'Socrates', I cannot distinguish between sense and *Bedeutung*; I see only the idea, which is psychological, and the object. (Russell 1904a)

Elsewhere, in his 1904 'Meinong's Theory of Complexes and Assumptions', he contends:

> [For Meinong,] a presentation, with the three elements of act, content and object, is taken to be wholly psychical, and the object, like the other two, is supposed to be part of the total mental state—so at least some passages suggest … As against this view, I should prefer to advocate what is, presumably, the distinguishing feature of a common-sense philosophy, namely, that the object of a presentation is *the actual external object itself*, and not any part of the presentation at all. (Russell 1904b, p. 441, emphasis added)[33]

---

concerning descriptive phrases which lack denotation: it either makes propositions that intuitively have truth-values truth-valueless, or it ushers in Meinongian objects (and the contradictions Russell thinks they generate). But nowhere in KAKD does he say anything which suggests that he intends to solve Frege's Puzzle *in full generality* or that the special epistemic features of acquaintance are part of the solution.

[33] Russell goes on to say 'Thus to take the case of the steeple: we have (1) external perception, having as its object the actual steeple itself, or rather the existence of the steeple, the wholly extra-mental material thing; (2) purely internal perception, having as its object the existence of either the act or the content of the previous external perception; (3) the perception of the object of higher order, consisting of the perception and the existence of the steeple combined by the relation of cognition' (Russell 1904b, p. 441).

 © Wishon 2016

Thus, just one year prior to the publication of OD, Russell has not yet restricted our acquaintance to sense-data. And in my opinion, there are no compelling reasons for thinking he changes his mind in the interim. In fact, to my knowledge, Russell never mentions sense-data as the direct objects of sensation until 1910 in the chapter 'On the Nature of Truth and Falsehood' in *Philosophical Essays* (hereafter *PE*)—well after the publication of OD. [34]

Granted, Russell undeniably says in OD that we are not acquainted with such things as 'the centre of mass of the solar system', 'other people's minds', or 'matter (*in the sense in which matter occurs in physics*)' (pp. 479–80, 492–3, emphasis added). Nevertheless, we need to pay careful attention to the qualification Russell makes with regard to matter. While it admittedly is open to multiple interpretations, one straightforward and plausible reading is simply that we do not have acquaintance with the fundamental building blocks (atoms, molecules, electrons, etc.) that our physical theories tell us make up familiar macroscopic physical objects.[35] If this is correct, then it is

[34] It is also worth noting that G.E. Moore first presents something like sense-data as the direct objects of perception only later in 1905 when he read his 'The Nature and Reality of Objects of Perception' on December 18 at the twenty-seventh meeting of the Aristotelian Society. See Hatfield (2013) for an informative discussion of Russell's and Moore's early views about sense-data and perception.

[35] Proops (2015) argues that similarities between Russell's wording in OD and remarks in his unpublished 1912 'On Matter' claiming that we are not acquainted with 'matter as it is dealt with by physical science' suggest that he already excludes our acquaintance with macroscopic physical objects in the earlier work. Proops further contends that since Russell includes the moon in the category of matter dealt with by physical science in the later paper, it should be included in the same category in the earlier one lest we be obliged 'to conclude that Russell's view of the subject matter of physics must have itself undergone an unexplained and fairly dramatic shift between 1905 and 1912' (2015, p. 56). In Proops' view, the qualification in OD instead plausibly marks Russell's intention to leave open the possibility that we are acquainted with the void-denying philosopher's 'subtle matter'.

It could be that Proops is correct. However, it is worth noting that Russell's only example of the matter with which we lack acquaintance in OD is precisely a material particle (p. 427). Moreover, in his 1903 *Principles of Mathematics* (hereafter *POM*), Russell suggests that, strictly speaking, the matter of physics is a collection of material particles, in various configurations, which behave in accord with the principles of rational dynamics.

To my knowledge, these discussions are Russell's last word on the topic prior to OD and Proops owes an explanation for why Russell's views about the subject matter of physics undergo a fairly dramatic shift between 1903 and 1905.

Regarding my own reading, it could be that Russell is simply speaking loosely in 'On Matter' when he says that the moon is a piece of matter, or more likely, his views about matter do in fact change between 1905 and 1912. After all, in *MPD* Russell himself reports 'When, however, after 1910, I had done all that I intended to do as regards pure mathematics, I began to think about the physical world' (p. 12), which also coincided with his shifting interests towards epistemology. Finally, Russell's adoption of structuralism about

 © Wishon 2016

anachronistic to understand Russell here as making the radical claim (characteristic of his post-1910 writings) that we lack acquaintance with any part of material reality rather than as offering a commonplace illustration of the importance of our ability to think and talk about things which elude direct sensory experience.

Furthermore, there is convincing evidence elsewhere in OD that Russell indeed allows for our acquaintance with ordinary, medium sized material objects (such as human bodies) and their sensible features. In particular, he unequivocally says that we can literally *see* other people (though not their minds) when considering the case of George IV:

> When we say, 'George IV wished to know whether Scott was the author of *Waverley*', we normally mean 'George IV wished to know whether one and only one man wrote *Waverley* and Scott was that man'; but we *may* also mean 'One and only one man wrote *Waverley*, and George IV wished to know whether Scott was that man'. In the latter, 'the author of *Waverley*' has *primary* occurrence; in the former, a *secondary*. The latter might be expressed by 'George IV wished to know, concerning the man who in fact wrote *Waverley*, whether he was Scott'. *This would be true, for example, if George IV had seen Scott at a distance, and had asked 'Is that Scott?'* (OD, p. 489, emphasis added)

While Russell is here ostensibly discussing differences in the scope descriptive phrases might take in giving belief reports, it has significant import beyond such issues when we consider the meaning of George IV's beliefs and utterances themselves, rather than reports about them. For when George IV wonders (aloud or to himself) 'Is that Scott?' he seems precisely to be wondering whether an (informative) identity of the genuine logical form of '$a = b$' is true or false. Nevertheless, George IV is manifestly unable to determine whether or not the object of his occurrent visual acquaintance is Scott. So if it turns out that George IV *is* in fact seeing Scott in the distance (which is a live possibility), we not only have a clear instance of Frege's Puzzle, we also have a decisive counterexample to the received view.[36]

---

physics in *POP* opens up the possibilities that for all we know 'matter as described by physics' is a collection of material particles, ideas, monads, qualities, etc., which marks a clear divergence from his earlier views about the subject matter of physics.

[36] I recently discovered that Kripke (2005) has come to the same conclusions about OD. On page 1,024, he remarks, 'In [OD], Russell introduces his famous distinction between knowledge by acquaintance and knowledge by description and says we often use definite descriptions to denote objects we know only by description and are not acquainted with. In particular, we're not acquainted with the centre of the mass of the sun. We are also not

 © Wishon 2016

But then what are we to make of Russell's remarks about 'Scott is Sir Walter' being a tautology? Do they not show that he *has* to accept something like the received view? I think there is good reason to be sceptical of this. There is a significant interpretive question about just how exactly Russell understands the notion of a 'tautology', and whether he has more than one phenomenon in mind. In particular, it is not at all clear that he thinks that all tautologies wear their logical status on their sleeves. To be sure, there are passages in which Russell seems to think that all tautologies can be known to be true a priori. For instance, in his 1920 'The Wisdom of Our Ancestors', he says:

> The proposition 'If Socrates is a man, then Socrate est un homme', can be known without studying Socrates, by merely knowing the English and French languages. In logical propositions it is the syntax, not the vocabulary, that is translated, but it remains the case that only an understanding of language is necessary in order to know a proposition of logic. (p. 405)

And immediately before that, he asserts that:

> As for logic and the so-called 'Laws of Thought', they are concerned with symbols; they give different ways of saying the same thing. It is for this reason that the truth of logical propositions can be known without studying the things to which they apply. (p. 405)

So it certainly appears that Russell thinks that tautologies should be transparent to a subject simply provided that he or she understands the terms involved (and they do not require too much technical skill to work out).[37]

However, we should be careful here. For it is arguable that Russell is here imagining someone who knows that 'Socrates is a man' means the same thing as 'Socrate est un homme' because he or she knows which expression in English should be translated into a given expression in French, and vice versa. In this sort of case, it is quite plausible that the bilingual subject would recognize the tautology for what it is, and likewise in the case in which the subject *already knows* that 'Scott' and 'Sir Walter' are two names for the same individual. But that is not at all what is going on in standard cases of misidentification. In these latter cases, the subject understands the meanings of each of the expressions involved, but manifestly does not know that they mean one

---

acquainted with *other minds*, he says at the end of the paper. However, there is no suggestion that Scott is to be identified with his mind here, nor that 'Scott' is anything other than a genuine proper name'. Also see footnote 43 on the same page.

[37] I am grateful to Kevin Klement for drawing my attention to these passages.

 © Wishon 2016

and the same thing. And there is good reason to think that Russell recognizes the difference here.[38]

We should also notice that in the passages from 'The Wisdom of Our Ancestors' Russell is claiming that *propositions of logic* can be known a priori. But it unclear that he takes tautologies and propositions of logic to amount to the same thing, even here.[39] In PLA, for instance, he says that a proposition of logic is 'somehow or other *like* a tautology' (p. 107, emphasis added). But that is not to equate them with each other; quite the contrary. Moreover, we should notice that it is a *necessary condition* of propositions of logic that they consist solely of variables. So the proposition 'for all x, $x = x$' is permitted to be a proposition of logic, but the proposition '$a = b$' is not, *precisely because it involves genuine referring expressions*. So it would be invalid to infer that Russell thinks that tautologies are knowable a priori from his claim that propositions of logic are.

One might object here that surely '$a = a$' is knowable a priori, and that '$a = b$', if true, is the very same proposition as '$a = a$', and so '$a = b$' must be knowable a priori as well. However, it is not clear at all that Russell would agree that '$a = a$' is a priori. For seemingly, he thinks that one must *derive* the knowledge that '$a = a$' from knowledge of the logical proposition 'for all x, $x = x$' *together with* knowledge that 'a' is a logically proper name, rather than a mere noise or truncated description.[40] In other words, one must also know something to the effect that 'there is an x, such that $x = a$'.[41] Yet he is quite clear that a

[38] See, for instance, the chapter on ostensive definition in *HK*.

[39] It is clear that Russell did not take propositions of logic to be tautologies in 1912 or prior. He came closer to such a view after 1914 under the influence of Wittgenstein. For example, see his remarks in *IMP* that 'all the propositions of logic have a characteristic which used to be expressed by saying that they were analytic, or that their contradictories were self-contradictory. This mode of statement, however, is not satisfactory ... Nevertheless, the characteristic of logical propositions that we are in search of is the one which was felt, and intended to be defined, by those who said that it consisted in deducibility from the law of contradiction. This characteristic ... for the moment, we may call *tautology*' (p. 206). However, as I will soon suggest, this just further goes to show that Russell is unsure of how exactly to understand the notion of 'tautology'.

[40] For instance, in *IMP* he says 'Suppose, for example, that ϕx is "always true"; let it be, say, the "law of identity", $x = x$. Then we may substitute for "x" any name we choose, and we shall obtain a true proposition. Assuming for the moment that "Socrates", "Plato", and "Aristotle" are names (a very rash assumption), *we can infer from the law of identity* that Socrates is Socrates, Plato is Plato, and Aristotle is Aristotle' (p. 178, emphasis added).

[41] In PLA he says, 'to understand a name you must be acquainted with the particular of which it is a name, *and you must know that it is the name of that particular*' (p. 67, emphasis added). While he goes on to say that understanding a name does not bring in the form of a

 © Wishon 2016

proposition of the form 'there is an x, such that $x = a$' is not a proposition of logic and can only be known to be true on empirical grounds (PLA, p. 107).

Consequently, it seems that for Russell, we can a priori *deduce* '$a = b$', if true, only if we combine the a priori knowledge that 'for all x, $x = x$' with antecedently possessed a posteriori knowledge that 'there is an x, such that $x = a$', that 'there is a y, such that $y = b$', and that '$x = y$'. If this is right, then the bilingual subject that knows that 'Socrates is a man' translates as 'Socrate est un homme' and the monolingual subject that already knows that 'Scott' and 'Sir Walter' are two names for the same person will recognize the relevant tautologies for what they are. But this is not so for the subject who sincerely wonders whether 'Scott is Sir Walter'.

Unfortunately, Russell's discussions of the nature of tautologies and their relationship to propositions of logic lack sufficient clarity to settle the matter one way or the other. I am inclined to think that the most likely reason for this is that his own thoughts on these issues are far from clear. Indeed, I think that we should take very seriously his admission in PLA that he is unable to say exactly what enables us to know logical propositions a priori (p. 108). Relatedly, we should pay equal attention to his following lamentation at the end of *IMP*:

> For the moment, I do not know how to define 'tautology'. It would be easy to offer a definition which might *seem* satisfactory for a while; but I know of none that I feel to be satisfactory, in spite of feeling thoroughly familiar with the characteristic of which a definition is wanted. At this point, therefore, for the moment, we reach the frontier of knowledge on our backward journey into the logical foundations of mathematics. (p. 208, emphasis in original)

While I am inclined to follow David Pears (1981) in suspecting that what Russell is really gesturing at is that identity statements of the genuine logical form of '$a = b$' are Kripkean a posteriori necessary identities, I fully acknowledge that the matter is far from clear (Kripke 1972). But this uncertainty cuts both ways. For it also follows that the defender of the received view of acquaintance cannot securely rest his or her case on Russell's remarks that 'Scott is Sir Walter' is a trivial tautology.

This leaves us with further interesting (but to my knowledge unexplored) interpretive questions about just how much interest Russell

---

proposition, what he seemingly has in mind is that, unlike in the case of predicates, it does not require understanding what it would be for a certain propositional function to take a range of values.

 © Wishon 2016

has in providing a general solution to Frege's Puzzle. He clearly has interest in failures of truth-preservation when co-designating genuinely referring expressions and descriptive phrases are substituted for one another, most notably in attitude ascriptions. But his principal interests in discussing Frege's sense–reference distinction in his works leading up to OD overwhelmingly concern the issues of expressions which lack corresponding referents and getting the truth-conditions of our thought and talk right, with an eye squarely on solving philosophical problems arising in logic and mathematics.[42] And it simply is not clear that he has *independent* interest in Frege's Puzzle, or in issues of informative identities involving genuine names.

In fact, in 'Letter to Frege' and *POM*, Russell shows a striking *ambivalence* towards Frege's Puzzle when genuinely referring expressions are involved. In 'Appendix A' of *POM*, for example, he casually remarks:

> [Frege's] theory of indication is more sweeping and general than mine, as appears from the fact that *every* proper name is supposed to have two sides [that is, sense and reference]. It seems to me that only such proper names as are derived from concepts by means of *the* can be said to have meaning, and that such words as *John* merely indicate without meaning. (p. 502, emphasis in original)

But ordinary names for persons such as 'John' are just the sorts of expressions for which Frege's Puzzle naturally arises, a point of which Russell is surely aware given that he is explicitly discussing 'Über Sinn and Bedeutung' (1892). So the fact that he makes no attempt to offer (or even gesture at the need for) a solution to the problem of informative identities regarding names is very suggestive. Unfortunately, a satisfactory investigation of these issues must wait until another occasion. But I hope that I have shown that whichever way it might turn out, it is doubtful that the received view of acquaintance will be vindicated.

7. Conclusion

I have argued that a number of recent Russell interpreters, such as Gareth Evans, Donald Davidson, John Campbell, and Ian Proops have attributed to Russell what I call 'the received view of acquaintance': the view that acquaintance safeguards us from misidentifying the objects of our acquaintance. I have also contended that Russell's discussions of phenomenal continua cases show that he does not accept the

[42] See *POM* and a number of relevant unpublished papers from 1903-4, including Russell (1994a; 1994b; 1994c; 1994d; 1994e).

 © Wishon 2016

received view, or 'Russell's Principle' strongly interpreted. Indeed, I have argued, the possibility of misidentifying the objects of acquaintance should be quite unsurprising given underappreciated aspects of Russell's overall theory of knowledge and acquaintance. Finally, I have discussed the radical impact that Russell's *actual* views on acquaintance have for our understanding of his well-known George IV case in OD. In particular, I have argued that the George IV case is seemingly *not* intended to be a one-size-fits-all solution to Frege's Puzzle (and similar substitution problems), and hence does not provide any compelling support for the received view of acquaintance.

The importance of this finding is not merely a historical one, however. There is a good deal of recent work in the philosophy of mind, the philosophy of language, and epistemology which takes Russell's notion of acquaintance as a key starting point. But if I am correct in interpreting Russell's actual views on acquaintance, then much of this contemporary work rests on a misunderstanding. For one thing that is clear is that if we are to follow Russell's lead we must reject 'Russell's Principle', as strongly interpreted, wholesale. Is that plausible? Just how *far* can we depart from the strong readings of 'Russell's Principle'?[43]

References

Alter, Torin and Yujin Nagasawa (eds) 2015: *Consciousness in the Physical World: Perspectives on Russellian Monism*. Oxford: Oxford University Press.

Alter, Torin and Sven Walter (eds) 2007: *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: Oxford University Press.

Amijee, Fatema 2013: 'The Role of Attention in Russell's Theory of Knowledge'. *British Journal for the History of Philosophy*, 21, pp. 1,175–93.

Austin, David F. 1990: *What's the Meaning of 'This'?: A Puzzle about Demonstrative Belief*. Ithaca, NY: Cornell University Press.

Bach, Kent 2010: 'Getting a Thing into a Thought'. In Jeshion 2010a, pp. 39–63.

[43] I would like to thank John Perry, Krista Lawlor, Mark Crimmins, Dagfinn Føllesdal, David Beisecker, John Campbell, Ian Proops, Russell Wahl, Bernard Linsky, Kevin Klement, Peter Hylton, Thomas Baldwin, Fatema Amijee, Robert Barnard, Timothy Yenter, Bryan Harper, Ralf Bader, and the anonymous referees at *Mind*. I am also grateful to audiences at Stanford University and at the 2011 Annual Meeting of the Bertrand Russell Society.

 © Wishon 2016

Beany, Michael (ed.) 2013: The Oxford Handbook to the History of Analytic Philosophy. Oxford: Oxford University Press.
Bigelow, John and Robert Pargetter 2006: 'Re-Acquaintance with Qualia'. *The Australasian Journal of Philosophy*, 84, pp. 353–78.
Brandl, Johannes and Wolfgang Gombocz (eds) 1989: *The Mind of Donald Davidson*. Amsterdam: Rodopi.
Campbell, John 2009: 'Consciousness and Reference'. In McLaughlin, Beckermann, and Walter 2009, pp. 648–62.
Clark, Romane 1981: 'Acquaintance'. *Synthese*, 46, pp. 231–45.
Davidson, Donald 1989: 'What is Present to the Mind?'. In his 2001, pp. 53–67. Originally published in Brandl and Gombocz 1989.
——2001: *Subjective, Intersubjective, Objective*. New York: Oxford University Press.
Davidson, Donald and Gilbert Harman (eds) 1972: *Semantics of Natural Language*. Dordrecht: Reidel.
Evans, Gareth 1983: *The Varieties of Reference*. New York: Oxford University Press.
Frege, Gottlob 1892: 'Über Sinn und Bedeutung', tr. Herbert Feigl. In *Martinich 2000, pp. 199–211. Originally published in Zeitschrift für Philosophie und philosophische Kritik*, 100, pp. 25–50.
——1980: *Philosophical and Mathematical Correspondence*, tr. Hans Kaal, Brian McGuinness (ed.). Chicago: University of Chicago Press.
Grote, John 1865: *Exploratio Philosophica: Rough notes on modern intellectual science, Part 1*. Cambridge: Deighton Bell.
Hale, Bob and Crispin Wright (eds) 1997: *A Companion to the Philosophy of Language*. Oxford: Blackwell.
Hatfield, Gary 2013: 'Perception and Sense-Data'. In Beany 2013, pp. 948–74.
James, William 1890: *The Principles of Psychology*. New York: Henry Holt.
Jeshion, Robin (ed.) 2010a: *New Essays on Singular Thought*. Oxford: Oxford University Press.
——(ed.) 2010b: 'Singular Thought: Acquaintance, Semantic Instrumentalism, and Cognitivism'. In Jeshion 2010a, pp. 105–40.
Johnston, Mark 1992: 'How to Speak of Colours'. *Philosophical Studies*, 68, pp. 221–63.
Kaplan, David 1978: 'On the Logic of Demonstratives'. *Journal of Philosophical Logic*, 8, pp. 81–98
——2005: 'Reading "On Denoting" on its Centenary'. *Mind*, 114, pp. 933–1,003.

© Wishon 2016

Klement, Kevin 2004: 'The Origins of the Propositional Functions Version of Russell's Paradox'. *Russell: The Journal of Bertrand Russell Studies*, 24, pp. 101–32.
Kripke, Saul 1972: 'Naming and Necessity'. In Davidson and Harman 1972, pp. 253–355, 763–9.
——1979: 'A Puzzle about Belief'. In Margalit 1979, pp. 239–83.
——2005: 'Russell's Notion of Scope'. *Mind*, 114, pp. 1,005–37.
Linsky, Bernard 2015: 'Acquaintance and Certainty in *The Problems of Philosophy*'. In Wishon and Linsky 2015, pp. 65–85.
Linsky, Bernard and Donovan Wishon (eds) forthcoming: *Acquaintance, Knowledge, and Logic: New Essays on Bertrand Russell's The Problems of Philosophy*. Stanford, CA: CSLI Publications.
Margalit, Avishai (ed.) 1979: *Meaning and Use*. Dordrecht: Reidel.
Martinich, A.P. (ed.) 2000: *The Philosophy of Language*, fourth edition. New York: Oxford University Press.
McLaughlin, Brian, Ansgar Beckermann ,and Sven Walter (eds) 2009: *The Oxford Handbook to Philosophy of Mind*. Oxford: Oxford University Press.
Moore, G. E. 1903: 'The Refutation of Idealism'. *Mind*, 12, pp. 433–53.
——1905-6: 'The Nature and Reality of Objects of Perception', *Proceedings of the Aristotelian Society*, 6, pp. 68–127.
Neale, Stephen 2005: 'A Century Later'. *Mind*, 114, pp. 809–71.
Pears, David 1981: 'The Function of Acquaintance in Russell's Philosophy'. *Synthese*, 46, pp. 149–66.
Perkins, Ray, Jr. 2007: 'Why "On Denoting"?'. *Russell: The Journal of Bertrand Russell Studies*, 27, pp. 24–40.
Perry, John 2001: *Reference and Reflexivity*. Stanford, CA: CSLI Publications.
Proops, Ian 2011: 'Russell on Substitutivity and the Abandonment of Propositions'. *The Philosophical Review*, 120, pp. 151–205.
——2014: 'Russellian Acquaintance Revisited'. *Journal of the History of Philosophy*, 52, pp. 779–811.
——2015: 'Certainty, Error, and Acquaintance in *The Problems of Philosophy*'. In Wishon and Linsky 2015, pp. 45–63.
Recanati, François 2010: 'Singular Thought: In Defence of Acquaintance'. In Jeshion 2010a, pp. 141–90.
Richard, Mark 1997: 'Propositional Attitudes'. In Hale and Wright, pp. 197–226.
Russell, Bertrand 1903: *The Principles of Mathematics*. London: Bradford and Dickens, rprt 1937.

 © Wishon 2016

——1904a: 'Letter to Frege'. In Frege 1980, pp. 166–70.
——1904b: 'Meinong's Theory of Complexes and Assumptions'. In Urquhart 1994, pp. 431–74. Originally published in *Mind*, 13, pp. 204–19.
——1905: 'On Denoting'. In *Urquhart 1994*, pp. 414–27. Originally published in Mind, 14, pp. 479–93.
——1910: *Philosophical Essays*. London: Longmans, Green, and Co.
——1911: 'Knowledge by Acquaintance and Knowledge by Description'. In Slater 1992, pp. 147–61. Originally published in *Proceedings of the Aristotelian Society*, 11, pp. 108–28.
——1912: *The Problems of Philosophy*. London: Williams and Norgate.
——1914: *Our Knowledge of the External World as a Field for Scientific Method in Philosophy*. Chicago: Open Court.
——1918: 'The Philosophy of Logical Atomism'. In Slater 1986, pp. 157–246. Originally published in *Monist*, 28, pp. 495–527 and *Monist*, 29, pp. 33–63, 190–222, and 345–80
——1919a: *Introduction to Mathematical Philosophy*. London: George Allen and Unwin.
——1919b: 'On Propositions: What They Are and How They Mean'. In Slater 1986, pp. 276–308. Originally Published in *Proceedings of the Aristotelian Society, Supplementary Volume: Problems of Science and Philosophy*, 2, pp. 1–43.
——1920: 'The Wisdom of Our Ancestors'. In Slater 1988, pp. 403–6. Originally published in *The Athenaeum*, 4,680, p. 43.
——1921: *The Analysis of Mind*. London: George Allen and Unwin.
——1923: 'Vagueness'. In Slater 1988, pp. 145–54. Originally published in *The Australasian Journal of Psychology and Philosophy*, 1, pp. 84–92.
——1927: *The Analysis of Matter*. London: Kegan Paul.
——1940: *An Inquiry into Meaning and Truth*. New York: W. W. Norton.
——1948: *Human Knowledge: Its Scope and Limits*. London: George Allen and Unwin.
——1957: 'Mr. Strawson on Referring'. In Slater 1997, pp. 630–4. Originally published in *Mind*, 66, pp. 385–9.
——1959: *My Philosophical Development*. New York: Simon and Schuster.
——1992a: *Theory of Knowledge: The 1913 Manuscript*. New York: Routledge.
——1992b: 'On Matter'. In Slater 1992, pp. 80–95.

 © Wishon 2016

——1994a: 'On the Meaning and Denotation of Phrases'. In Urquhart 1994, pp. 286–96.
——1994b: 'Dependent Variables and Denotation'. In Urquhart, pp. 297–304.
——1994c: 'Points about Denoting'. In Urquhart, pp. 305–13.
——1994d: 'On Meaning and Denotation'. In Urquhart, pp. 314–58.
——1994e: 'On Fundamentals'. In Urquhart, pp. 359–413. Russell, Bertrand and Alfred North Whitehead 1910: *Principia Mathematica*, Volume 1. Cambridge: Cambridge University Press.
Salmon, Nathan 1986: *Frege's Puzzle*. Atacadero, CA: Ridgeview Press.
Slater, John G. (ed.) 1986: *The Collected Papers of Bertrand Russell, Vol. 8, The Philosophy of Logical Atomism and Other Essays, 1914-19*. London: George Allen and Unwin.
Slater, John G. (ed.) 1988: *The Collected Papers of Bertrand Russell, Vol. 9: Language, Mind and Matter, 1919-26*. Boston: Unwin Hyman.
Slater, John G. (ed.) 1992: *The Collected Papers of Bertrand Russell, Vol. 6: Logical and Philosophical Papers, 1909-13*. New York: Routledge.
Slater, John G. (ed.) 1997: *The Collected Papers of Bertrand Russell, Vol. 11: Last Philosophical Testament, 1943-68*. New York: Routledge.
Soames, Scott 2010: *The Philosophy of Language*. Princeton: Princeton University Press.
Strawson, Peter F. 1950: 'On Referring'. Mind, 59, pp. 320–44.
Sullivan, Arthur 2010: 'Millian Externalism'. In Jeshion 2010a, pp. 246–70.
Szabó, Zoltan 2005: 'The Loss of Uniqueness'. *Mind*, 114, pp. 1185–222.
Urquhart, Alisdair (ed.) 1994: *The Collected Papers of Bertrand Russell, Vol. IV: Foundations of Logic, 1903-5*. New York: Routledge.
Wahl, Russell 2007: '"On Denoting" and the Principle of Acquaintance'. *Russell: The Journal of Bertrand Russell Studies*, 27, pp. 7–23.
——2015: 'Sense-Data and the Inference to Material Objects: The Epistemological Project in *Problems* and its Fate in Russell's Later Work'. In Wishon and Linsky 2015, pp. 87–105.
White, Stephen L. 2007. 'Property Dualism, Phenomenal Concepts, and the Semantic Premise'. In Alter and Walter 2007, pp. 210–48.

 © Wishon 2016

Williamson, Timothy 2000: *Knowledge and Its Limits*. Oxford: Oxford University Press.
Wishon, Donovan 2015: 'Russell on Russellian Monism'. In Alter and Nagasawa 2015, pp. 91–118.
Wishon, Donovan and Bernard Linsky (eds) 2015: *Acquaintance, Knowledge, and Logic: New Essays on Bertrand Russell's The Problems of Philosophy*. Stanford, CA: CSLI Publications.

© Wishon 2016

Maximally Consistent Sets of Instances of Naive Comprehension

Luca Incurvati
University of Amsterdam and Institute for Logic, Language and Computation
l.incurvati@uva.nl

Julien Murzi
University of Salzburg and Munich Center for Mathematical Philosophy, Ludwig Maximilian University
julien.murzi@sbg.ac.at

Paul Horwich (1990) once suggested restricting the T-schema to the maximally consistent set of its instances. But Vann McGee (1992) proved that there are multiple incompatible such sets, none of which, given minimal assumptions, is recursively axiomatizable. The analogous view for set theory—that Naive Comprehension should be restricted according to consistency maxims—has recently been defended by Laurence Goldstein (2006, 2013). It can be traced back to W. V. Quine (1951), who held that Naive Comprehension embodies the only really intuitive conception of set and should be restricted as little as possible. The view might even have been held by Ernst Zermelo (1908), who, according to Penelope Maddy (1988), subscribed to a 'one step back from disaster' rule of thumb: if a natural principle leads to contradiction, the principle should be weakened just enough to block the contradiction. We prove a generalization of McGee's Theorem, and use it to show that the situation for set theory is the same as that for truth: there are multiple incompatible sets of instances of Naive Comprehension, none of which, given minimal assumptions, is recursively axiomatizable. This shows that the view adumbrated by Goldstein, Quine, and perhaps Zermelo, is untenable.

Naive set theory is underwritten by the *Naive Comprehension Schema*, which consists of all instances of

$$\text{(NC)} \quad \exists y \forall x (x \in y \leftrightarrow \Phi(x))$$

where $\Phi(x)$ is any formula in $\mathcal{L}_{\in}$ (the language obtained by adding '$\in$' to the language of first-order logic with identity) in which x is free and which contains no free occurrences of y. The schema is intended to express the thought that every condition determines a set: the set of all and only those objects to which the condition applies.

Mind, Vol. 126 . 502 . April 2017 **© Incurvati and Murzi 2017**
doi:10.1093/mind/fzv192 Advance Access publication 19 January 2017

Naive truth theories adhere to the unrestricted *T-schema*:

(T) $\mathsf{Tr}(\ulcorner\phi\urcorner) \leftrightarrow \phi$

where Tr is a truth predicate and $\ulcorner\phi\urcorner$ is, for present purposes, the numerical code of ϕ relative to some fixed coding scheme. The principle trades on the putative fact that ϕ and '"ϕ" is true' have the same content, and hence should be at least materially equivalent.

Naivety has a price, however: unless classical logic is quite radically weakened, paradox looms. Let $\Phi(x)$ in (NC) be $x \notin x$, and you get Russell's Paradox. Take ϕ in (T) to be $\lambda = \neg\mathsf{Tr}(\ulcorner\lambda\urcorner)$, and you get the Liar Paradox. This much is well known. What is not so well known is whether (NC) and (T) can be consistently restricted in a way that is faithful to the spirit of the naive conceptions they characterize.

In this note, we examine the view that naive principles should be restricted according to consistency maxims. A proposal along these lines was once made by Paul Horwich (1990) for truth, and has recently been defended by Laurence Goldstein (2006, 2013) for both truth and set theory.

For the case of sets, the view goes back at least as far as W. V. Quine (1951). In his view, (NC) embodies the only really intuitive conception of set, and should, in keeping with the maxim of minimal mutilation, be restricted as little as possible so as to avoid the paradoxes. He writes:

> Only because of Russell's paradox and the like do we not adhere to the naive and unrestricted comprehension schema ... Having to cut back because of the paradoxes, we are well advised to mutilate no more than what may fairly be seen as responsible for the paradoxes. (Quine 1951, pp. 50–1)

The view might even have been held at one time by Ernst Zermelo (1908, pp. 263–4). According to Penelope Maddy (1988, p. 485), Zermelo saw himself as giving us as much of (NC) as possible without inconsistency. With Zermelo, Maddy suggests, emerged the rule of thumb she labels 'one step back from disaster': if a natural principle leads to contradiction, the principle should be weakened just enough to block the contradiction.

Strictly speaking, Goldstein also requires acceptable instances of the naive schemata not to entail tautologies, as well as contradictions.[1]

[1] Tautologies and contradictions, says Goldstein, lack content, and hence 'fail to state conditions or to give a recipe for determining whether any given object belongs in the set' (Goldstein 2013, p. 35) in the case of sets, and fail to specify coherent truth conditions in

 © Incurvati and Murzi 2017

However, it is easy to see that such a proposal would involve quite a radical departure from classical logic.[2] We are here interested in the prospects for curtailing the Naive Comprehension Schema whilst retaining classical logic, and focus on the idea that the problematic instances to be banned are the inconsistent ones—the ones giving rise to paradox.[3] We argue that the prospects for such a strategy are not good: restrictions of the kind Goldstein advocates fail for essentially the same reasons that, as Vann McGee (1992) first noticed, Horwich's proposed restriction of the T-schema also fails.

1. Consistency

Goldstein writes:

> The naive comprehension axiom captures a very clear conception of sets, and it makes sense to tamper with it as little as possible, identifying and justifying the restriction that must be made to its application, rather than building up set theory from entirely new foundations. (Goldstein 2013, pp. 35–6)

So how can (NC) be made consistent, without tampering too much with the naive conception of set?

If classical logic is to be preserved, Goldstein's idea amounts to restricting (NC) according to the following injunction:

Consistent: An instance ϕ of (NC) is admissible just in case $\phi \nvdash \bot$.[4]

---

the case of truth. Thus, $\exists y \forall x(x \in y \leftrightarrow x \notin x)$ is not an acceptable instance of (NC), and $\mathsf{Tr}(\ulcorner \lambda \urcorner) \leftrightarrow \lambda$ is not an acceptable instance of (T): both entail a contradiction, and hence are 'vacuous' and specify conditions that are 'impossible to satisfy' (Goldstein 2006, p. 886). In Goldstein's view, this is the reason for thinking that the set-theoretic and semantic paradoxes belong to the same family: 'Herein lies the similarity between these two families of paradox. The sets and the statements we thought we were dealing with are irredeemably ill-defined, and so do not exist' (Goldstein 2006, p. 887). It is then only natural, Goldstein suggests, to restrict (NC) and (T) to instances that are neither vacuous nor impossible to satisfy. The restriction, he adds, is a principled one: it follows from the Tractarian view that logical truths and falsehoods do not express propositions (Wittgenstein 1922).

[2] In classical logic, tautologies are entailed by any sentences whatsoever, and hence also by instances of (NC). Goldstein is, of course, well aware of this, and in fact endorses a non-classical logic (see Goldstein 2012).

[3] To be sure, a different strategy for the naive set theorist would be not to restrict (NC), and, in order to avoid triviality, *weaken the logic* instead. See, for example, Grishin (1981), Weber (2010, 2012). In the remainder of the paper, we focus on *classical* naive but consistent set theory. We briefly return to non-classical naive set theory in footnote 15 below.

[4] Here and throughout, '$\vdash$' stands for first-order deducibility.

 © Incurvati and Murzi 2017

Unfortunately, this restriction will not do. The problem is that it is easy to exhibit instances of (NC) which are consistent given extremely weak assumptions—and therefore admissible according to *Consistent*—but together lead to inconsistency. We give two examples.

The first example rests on the observation that instances of (NC) can be obtained by considering formulae $\Phi(x)$ containing free variables other than x, known as *parameters*. Such instances are assumed to be prefaced by universal quantifiers binding these free variables. One such instance—in effect, also an instance of ZFC's Separation Schema—asserts, for any set, the existence of the subset of all non-self-membered members of that set. Trouble arises when we consider this instance in conjunction with one asserting the existence of a universal set.

Example 1. The following are instances of (NC):

(1) $\exists y \forall x(x \in y \leftrightarrow x = x)$

(2) $\forall z \exists y \forall x(x \in y \leftrightarrow (x \in z \ \wedge \ x \notin x))$

(1) has a model consisting of just one object a such that $a = \{a\}$.[5] (2) is true in any model consisting of a single hereditarily finite set.[6] The simplest such model is the one consisting of just one object a such that $a \notin a$, that is, the model consisting of just the empty set—the first hereditarily finite set.[7] Hence, (1) and (2) are both admissible according to *Consistent*. But it is routine to derive a contradiction from (1) and (2) together.[8]

[5] Alternatively, if one prefers a model without non-well-founded sets, one can consider a model which consists of a looped one-node graph, in which variables range over nodes, and in which $x \in y$ just in case there is a directed edge from the node assigned to y to the node assigned to x.

[6] A set is said to be *hereditarily finite* iff it is finite, its members are finite, the members of its members are finite, and so on. In other words, a set is hereditarily finite iff its transitive closure is finite.

[7] In fact, as a referee pointed out, it is clear that (2) holds in any model satisfying $\forall x(x \notin x)$. Take any set z in such a model and let y be z itself. Then (2) reduces to $\forall x(x \in z \leftrightarrow (x \in z \ \wedge \ x \notin x))$. But since $\forall x(x \notin x)$ holds in this model, this reduces to the logical truth $\forall x(x \in z \leftrightarrow x \in z)$.

[8] (1) immediately entails $\forall x(x \in a \leftrightarrow x = x)$. Instantiating (2) with a (the universal set) we have that there is a set of all non-self-membered members of a: $\exists y \forall x(x \in y \leftrightarrow (x \in a \ \wedge \ x \notin x))$. From this, we get $b \in b \leftrightarrow (b \in a \ \wedge \ b \notin b)$. But instantiating $\forall x(x \in a \leftrightarrow x = x)$ with b we also have that $b \in a \leftrightarrow b = b$. Therefore, $b \in b \leftrightarrow b \notin b$. Contradiction.

© Incurvati and Murzi 2017

In light of Example 1, one might be tempted to ban instances of (NC) that are obtained using parameters. Our second example shows that this is not enough to dispense with the problem. The example makes use of the following observation, which goes back to George Boolos (1993, p. 233):

Lemma 2. For each ϕ in $\mathcal{L}_{\in}$,

(i) $\vdash \exists y \forall x(x \in y \leftrightarrow (\neg\phi \wedge x \notin x)) \rightarrow \phi$

(ii) $\exists y \forall x(x \notin y) \vdash \phi \rightarrow \exists y \forall x(x \in y \leftrightarrow (\neg\phi \wedge x \notin x))$[9]

Example 3. Take your favourite σ such that both σ and $\neg\sigma$ are consistent—for instance, let σ be shorthand for 'There are at least thirty objects'. Then the following are instances of (NC):

(3) $\exists y \forall x(x \in y \leftrightarrow (\neg\sigma \wedge x \notin x))$

(4) $\exists y \forall x(x \in y \leftrightarrow (\sigma \wedge x \notin x))$

By Lemma 2(ii), any proof of a contradiction from (3) could be transformed, in the presence of the Empty Set Axiom, into a proof of a contradiction from σ, and similarly for (4). Hence, in the presence of the Empty Set Axiom, (3) and (4) are both admissible instances of (NC) according to *Consistent*. But by Lemma 2(i), (3) entails σ, and (4) entails $\neg\sigma$. Contradiction.

[9] For (i), suppose $\exists y \forall x(x \in y \leftrightarrow (\neg\phi \wedge x \notin x))$. Now suppose towards $\exists$-elimination that $\forall x(x \in a \leftrightarrow (\neg\phi \wedge x \notin x))$. Then we have that $a \in a \leftrightarrow (\neg\phi \wedge a \notin a)$. Now suppose $a \in a$. Then $\neg\phi \wedge a \notin a$ and so $a \notin a$. Contradiction. So $a \notin a$. But then suppose $\neg\phi$. Then $a \notin a \wedge \neg\phi$, and so $a \in a$. Contradiction again. By classical *reductio* it follows that ϕ. But ϕ has been obtained on the supposition that $\forall x(x \in a \leftrightarrow (\neg\phi \wedge x \notin x))$. Hence, by $\exists$-elimination we have ϕ on the overall supposition that $\exists y \forall x(x \in y \leftrightarrow (\neg\phi \wedge x \notin x))$. The result follows by conditional proof.

For (ii), first notice that $b \notin a$ follows from our premiss $\exists y \forall x(x \notin y)$ and suppose ϕ. Now suppose $b \in a$. This contradicts $b \notin a$, and so we can infer $\neg\phi \wedge b \notin b$. Next, suppose $\neg\phi \wedge b \notin b$, and further assume $b \notin a$ for $\neg$-introduction. Since both ϕ and $\neg\phi$ are among our assumptions, $b \in a$ follows by classical *reductio*. We have derived $\neg\phi \wedge b \notin b$ from $b \in a$ and $b \in a$ from $\neg\phi \wedge b \notin b$. By $\leftrightarrow$-introduction, it follows that $b \in a \leftrightarrow (\neg\phi \wedge b \notin b)$. Applying $\forall$-introduction and $\exists$-introduction, we obtain $\exists y \forall x(x \in y \leftrightarrow (\neg\phi \wedge x \notin x))$. Thus, we have $\exists y \forall x(x \in y \leftrightarrow (\neg\phi \wedge x \notin x))$ on the overall supposition that ϕ. The result follows by conditional proof.

 © Incurvati and Murzi 2017

In short: instances of (NC) that are admissible according to *Consistent* may be *jointly* inconsistent, and hence unacceptable.[10] This suggests that naive but consistent set theorists must focus on acceptable *sets* of instances of (NC)—a suggestion to which we now turn.

2. Maximal consistency

The naive set theorist might take the results of the previous section to simply show that, whenever the addition of a consistent instance of (NC) yields an inconsistency, the instance is not acceptable after all. That is, she might take the naive principles to be so intuitive that we should keep as many instances of them as we consistently can. More formally, the acceptable instances of (NC) must form a *maximally consistent* set:

> *Maximally Consistent*: A set of instances Γ of (NC) is admissible just in case $\Gamma \nvdash \bot$ and if Λ is a set of instances of (NC) and properly extends Γ, then $\Lambda \vdash \bot$.[11]

In fact, this proposal seems to better capture the spirit of Quine's remark quoted above, and seems more in keeping with the 'one step back from disaster' rule of thumb, which, if Maddy is right, the early Zermelo subscribed to. The approach was once advocated by Horwich for truth:

> [I]t suffices for us to concede that certain instances of [the T-schema (T)] are not to be included as axioms of the minimal theory, and to

[10] The situation here is analogous to that arising in the context of the debate on the acceptability of *abstraction principles*. These are principles of the form

$$\S(\alpha) = \S(\beta) \leftrightarrow \alpha \approx \beta$$

where § is an operator taking items of type α and β to be objects, and $\approx$ is an equivalence relation on items of that type. In the *Grundlagen*, Frege (1884) considered and rejected the possibility of using a principle of this sort to contextually define the concept of number. And he later took extensions to be governed by an abstraction principle stating that two concepts have the same extension just in case they are coextensive. As is well known, the principle concerning extensions is inconsistent, so any attempt to defend the legitimacy of using abstraction principles to define mathematical concepts (such as Wright 1983) must address the question of which abstraction principles are acceptable to this end. But—and here lies the analogy with the case of (NC)—it will not do simply to say that the acceptable abstraction principles are the consistent ones. For there are abstraction principles that are individually consistent but jointly inconsistent (see, for example, Boolos 1990). For more on the acceptability of abstraction principles, see Fine (2002).

[11] As usual, we say that a set Λ *extends* a set Γ iff Γ is a subset of Λ, and that it *properly extends* Γ iff Γ is a proper subset of Λ.

 © Incurvati and Murzi 2017

note that the principles governing our selection of excluded instances are, in order of priority: (a) that the minimal theory not engender 'liar-type' contradictions; (b) that the set of excluded instances be as small as possible; and … (c) that there be a constructive specification of the excluded instances that is as simple as possible. (Horwich 1990, p. 42)

The view faces a twofold problem, however, as McGee (1992) first pointed out. First, in the presence of a theory strong enough to interpret Robinson's Arithmetic (henceforth Q), maximally consistent sets of instances of (T) are not recursively axiomatizable, and hence it is unclear whether they can be said to be *theories* of truth. Second, there exist several maximally consistent, and yet incompatible, sets of instances of (T). We are effectively presented with an embarrassment of riches, and it would seem arbitrary to choose one over another. We now show that the suggestion that (NC) must be restricted according to *Maximally Consistent* faces exactly the same difficulties.

3. Generalizing McGee's result

We want to show that McGee's result admits of a natural generalization. To this end, we first introduce some notation and terminology. If T is a theory, we write '$\Gamma \vdash_{\mathsf{T}} \phi$' for '$\Gamma \cup \mathsf{T} \vdash \phi$'. And we make use of the following definitions:

Definition 4. Let Γ be a set of sentences and Σ a first-order schema.

(i) Γ is T-*consistent* iff $\Gamma \nvdash_{\mathsf{T}} \bot$.
(ii) Γ is Σ-*maximally* T-*consistent* iff Γ is T-consistent and any set which properly extends Γ and contains Σ-instances not in Γ is T-inconsistent.
(iii) Γ is *maximally* T-*consistent* iff it is T-consistent and any set which properly extends Γ is T-inconsistent.

Clearly, this is an increasing order: if Γ is maximally T-consistent, then it is Σ-maximally T-consistent; and if Γ is Σ-maximally T-consistent, then it is T-consistent.

Now let S be a consistent first-order theory entailing the axioms of some theory S', let Σ be a first-order schema, and let $\mathcal{L}$ be a countable first-order language. Suppose we relatively interpret in $\mathcal{L}$ the language

 © Incurvati and Murzi 2017

obtained by adding to the language of S any non-logical expression in Σ. We can then prove the following:

Theorem 5. Suppose that for each $\phi \in \mathcal{L}$ there is an instance Σ_ϕ of Σ such that

$$\vdash_{S'} \phi \leftrightarrow \Sigma_\phi$$

Then:

(1) For any S-consistent set Δ of sentences in $\mathcal{L}$, there is a Σ-maximally S-consistent set Γ of Σ-instances such that $\Gamma \vdash_S \delta$ for every $\delta \in \Delta$.

(2) If Ξ is a Σ-maximally S-consistent set of Σ-instances, then, for every $\psi \in \mathcal{L}$, $\Xi \vdash_{S'} \psi$ or $\Xi \vdash_{S'} \neg\psi$.

Proof. For (1), let $\Theta = \{\Sigma_\delta : \delta \in \Delta\}$. Note that $\Theta \vdash_S \delta$ for every $\delta \in \Delta$, since

$$\vdash_S \delta \leftrightarrow \Sigma_\delta$$

for every $\delta \in \Delta$. Moreover, since Δ is S-consistent, so is Θ. Hence, by Zorn's Lemma, Θ can be extended to a Σ-maximally S-consistent set Γ of Σ-instances (see Lemma 8 in Appendix A for details).

For (2), let Ξ be a Σ-maximally S-consistent set of Σ-instances, and take a $\psi \in \mathcal{L}$. Since Ξ is S-consistent, then either $\Xi \cup \{\psi\}$ or $\Xi \cup \{\neg\psi\}$ is S-consistent.[12] So either $\Xi \cup \{\Sigma_\psi\}$ or $\Xi \cup \{\Sigma_{\neg\psi}\}$ is S-consistent. But since Ξ is Σ-maximally S-consistent, either $\Sigma_\psi \in \Xi$ or $\Sigma_{\neg\psi} \in \Xi$. Thus, $\Xi \vdash_{S'} \psi$ or $\Xi \vdash_{S'} \neg\psi$. □

Note that although the first part of the theorem only guarantees the existence of a certain Σ-maximally S-consistent set Γ of Σ-instances, it follows from the second part that a corresponding maximally S-consistent set can be obtained simply by closing Γ under $\vdash_S$.[13] Next, note that we could take S to be a theory which interprets Q.

[12] For suppose Ξ is S-consistent but neither $\Xi \cup \{\psi\}$ nor $\Xi \cup \{\neg\psi\}$ is S-consistent. From the fact that $\Xi \cup \{\psi\}$ is S-inconsistent it follows that $\Xi \cup \{\psi\} \vdash_S \neg\psi$ and so, by the deduction theorem, that $\Xi \vdash_S \psi \rightarrow \neg\psi$. But since $(\psi \rightarrow \neg\psi) \rightarrow \neg\psi$ is a first-order validity and hence $\Xi \vdash_S (\psi \rightarrow \neg\psi) \rightarrow \neg\psi$, we have that $\Xi \vdash_S \neg\psi$. Similar reasoning can be used to show that $\Xi \vdash_S \psi$ follows from the fact that $\Xi \cup \{\neg\psi\}$ is S-consistent. But then $\Xi \vdash_S \neg\psi$ and $\Xi \vdash_S \psi$, contradicting the initial supposition that Ξ is S-consistent.

[13] For let Γ be a Σ-maximally S-consistent set of Σ-instances and let $\mathrm{Ded}_S(\Gamma)$ be the set of all ϕ such that $\Gamma \vdash_S \phi$. Clearly, $\mathrm{Ded}_S(\Gamma)$ is S-consistent. Now suppose there is a set Λ which properly extends $\mathrm{Ded}_S(\Gamma)$. Then, there is some $\phi \in \mathcal{L}$ such that $\phi \notin \mathrm{Ded}_S(\Gamma)$ but $\phi \in \Lambda$. Since Γ is a Σ-maximally S-consistent set of Σ-instances, it follows from the second part of Theorem

 © Incurvati and Murzi 2017

It would then follow from the second part of the theorem that the theory obtained by adding any Σ-maximally S-consistent set of Σ-instances to S is negation complete, and hence non-recursively axiomatizable.[14] Finally, although we have used Zorn's Lemma in the proof of (1), this could be dispensed with by constructing Γ step by step (see Appendix A for details).

Given the above generalization of McGee's proof, we can obtain the desired result in the case that interests us by using Boolos's observation mentioned above. (McGee's result can be recovered as a special case of Theorem 5, as we show in Appendix B below.) Let R be some consistent theory entailing the axioms of the theory whose sole non-logical axiom is

$$(\mathsf{NC}_{\varnothing}) \quad \exists y \forall x (x \in y \leftrightarrow x \neq x)$$

And suppose that we have relatively interpreted the language of R in a countable first-order language $\mathcal{L}'$. We have:

Lemma 6. For each $\phi \in \mathcal{L}'$,

$$\vdash_{\mathsf{NC}_{\varnothing}} \phi \leftrightarrow \exists y \forall x (x \in y \leftrightarrow (\neg\phi \ \wedge \ x \notin x))$$

Proof. Immediate from Lemma 2. □

But together with Theorem 5, Lemma 6 immediately yields:

Corollary 7

(1) For any R-consistent set Δ of sentences in $\mathcal{L}'$, there is a (NC)-maximally R-consistent set Γ of (NC)-instances such that $\Gamma \vdash_{\mathsf{R}} \delta$ for every $\delta \in \Delta$.
(2) If Ξ is a (NC)-maximally R-consistent set of (NC)-instances, then, for every $\psi \in \mathcal{L}'$, $\Xi \vdash_{\mathsf{NC}_{\varnothing}} \psi$ or $\Xi \vdash_{\mathsf{NC}_{\varnothing}} \neg\psi$.

---

5 that for every $\psi \in \mathcal{L}$ either $\psi \in \mathrm{Ded}_{\mathsf{S}}(\Gamma)$ or $\psi \notin \mathrm{Ded}_{\mathsf{S}}(\Gamma)$. So $\neg\phi \in \mathrm{Ded}_{\mathsf{S}}(\Gamma)$. But since Λ is an extension of $\mathrm{Ded}_{\mathsf{S}}(\Gamma)$, this means that $\neg\phi \in \Lambda$ too. Thus, Λ is S-inconsistent.

[14] That such a theory cannot be recursively axiomatized follows from J. Barkley Rosser's (1936) strengthening of Gödel's (1931) First Incompleteness Theorem, which establishes that no consistent recursively axiomatizable theory interpreting Q is negation complete.

 © Incurvati and Murzi 2017

4. Maximal consistency is not enough

Corollary 7 has two key philosophical consequences. First, take any theory which interprets Q and contains a set of instances of (NC) which includes $\mathsf{NC}_{\varnothing}$ and is (NC)-maximally consistent with the rest of the theory. The second part of the corollary tells us that any such theory is negation complete, and hence cannot be recursively axiomatized. Clearly, however, any set theory worth its name should interpret Q. Moreover, $\mathsf{NC}_{\varnothing}$, which simply asserts the existence of the empty set, should be an acceptable instance of (NC) if anything is. Thus there is no reasonably strong recursively axiomatizable theory that satisfies the injunction dictated by *Maximally Consistent*: to identify our theory of sets with all instances of (NC) that are admissible according to *Maximally Consistent* would seem to fall short of providing a set *theory*.

Second, even setting this problem aside, the proposed emendation of naive set theory is fundamentally flawed. The first part of Corollary 7 gives us a recipe for constructing several admissible, and yet incompatible, sets of instances of (NC), each of which has an equal claim to embody the naive but consistent conception of set on offer.

To wit, take a σ such that σ and $\neg\sigma$ are (individually) consistent with the existence of the empty set. Then start with the sets $\{\exists y\forall x(x \in y \leftrightarrow (\neg\sigma \wedge x \notin x)), \mathsf{NC}_{\varnothing}\}$ and $\{\exists y\forall x(x \in y \leftrightarrow (\sigma \wedge x \notin x)), \mathsf{NC}_{\varnothing}\}$, and use Corollary 7 to extend each of these sets to a (NC)-maximally consistent set of (NC)-instances. The resulting two (NC)-maximally consistent sets are incompatible given Lemma 6: taking ϕ as σ in Lemma 6, we get the result that the first set implies σ, while taking ϕ as $\neg\sigma$ we have the result that the second set implies $\neg\sigma$. In order to restore consistency, the naive but consistent set theorist must settle on one such (NC)-maximally consistent set as embodying her conception. Yet we are given no guidance at all as to how this choice is to be made: consistency was meant to be our sole guide for restricting (NC).

To put matters in different terms, the naive but consistent set theorist adopts the 'one step back from disaster' rule of thumb, which urges us to restrict (NC) just enough to avoid inconsistency. However, there are several, incompatible directions in which such a step back could be taken. Absent some other rule of thumb, we are at a loss as to what direction should be favoured. We are inclined to interpret the provision of several mutually incompatible tamperings of the naive conception of set as the provision of no conception at all.

 © Incurvati and Murzi 2017

If our logic is to remain classical, or classical enough, yet more tinkering with (NC) is required.[15]

5. Concluding remarks

The semantic and set-theoretic paradoxes are standardly taken to show that the naive conceptions of set and truth are untenable, at least to the extent that we want to hold on to classical logic. Our results show that more is in fact true: such conceptions remain untenable even if, following the likes of Horwich and Quine, we give up precisely the portion of naivety that generates paradox. If our logic is to remain classical, a more radical departure from the naive schemata is required. This is the strategy pursued by, for instance, groundedness approaches (Leitgeb 2005) and revision theories in the case of truth (Gupta and Belnap 1993), and by the iterative conception (Boolos 1971) and the graph conception (Incurvati 2014) in the case of sets. Our results do not adjudicate between these approaches, and leave it open which of them, if any, should be favoured. Still, they do suggest that *some* approach of this kind is called for. Our final diagnosis is not different from McGee's: we take our results to be evidence that restrictions of the naive set-theoretic and semantic principles need to be guided by *non-naive* conceptions of sets and truth.[16]

Appendix A

We first prove the lemma (a version of Lindebaum's Lemma) used in the proof of part (1) of Theorem 5:

[15] Another option would be to both restrict (NC) according to consistency maxims *and* substantially weaken classical logic. We do not know whether our results would then carry over: they hold for classical logical consequence, but they need not continue to hold for weaker consequence relations. As mentioned above, Goldstein himself endorses a non-classical logic, and hence our results need not undermine his proposal. Here we simply notice that the suggestion of weakening the logic would be open to the potential problem that a set theory that is not strong enough to validate the results of this paper might be too weak. In any event, pending a precise description of the logic Goldstein favours, the issue remains highly speculative.

[16] Thanks to David Corfield, Robert van Rooij, Frank Veltman, Zach Weber and Jon Williamson for valuable discussion, and to Tim Button, Laurence Goldstein, Øystein Linnebo and three referees for this journal for extremely helpful comments on earlier versions of this material. Julien Murzi gratefully acknowledges the Alexander von Humboldt Foundation, the British Academy and the Leverhulme Trust for generous financial support during the time this paper was written. This paper is dedicated to the memory of Laurence Goldstein.

 © Incurvati and Murzi 2017

Lemma 8. Each S-consistent set of Σ-instances can be extended to a Σ-maximally S-consistent set of Σ-instances.

Proof. Suppose Ξ is an S-consistent set of Σ-instances. Let P be the set of all S-consistent sets of Σ-instances extending Ξ, partially ordered by inclusion $\subseteq$. If $\mathcal{C}$ is a chain of elements in P, then $\bigcup \mathcal{C}$ is S-consistent. So $\bigcup \mathcal{C}$ is in P, since $\bigcup \mathcal{C}$ contains every S-consistent set of Σ-instances extending Ξ. By Zorn's Lemma, P has a maximal element Γ. Now suppose that Γ is not Σ-maximally S-consistent. Then there is some $\psi \in \mathcal{L}$ such that $\Gamma \nvdash_S \Sigma_\psi$ and $\Gamma \nvdash_S \Sigma_{\neg\psi}$. If $\Gamma \nvdash_S \Sigma_\psi$ then $\Sigma_\psi \notin \Gamma$. And if $\Gamma \nvdash_S \Sigma_{\neg\psi}$, then $\Gamma \cup \{\Sigma_\psi\}$ is S-consistent and therefore in P. So $\Gamma \cup \{\Sigma_\psi\}$ is an S-consistent set of Σ-instances that properly extends Γ, contradicting the fact that Γ is a maximal element of P. Therefore, Γ is Σ-maximally S-consistent. □

We then prove part (1) of Theorem 5 without using Zorn's Lemma:

Proof. Enumerate the sentences of $\mathcal{L}$ as $\psi_0, \psi_1, \psi_2, \ldots$, and define an increasing sequence $\Gamma_0, \Gamma_1, \Gamma_2, \ldots$ of finite sets of Σ-instances as follows:

$$\Gamma_0 = \{\Sigma_\delta : \delta \in \Delta\}$$

$$\Gamma_{n+1} = \begin{cases} \Gamma_n \cup \{\Sigma_{\psi_n}\} & \text{if } \Gamma_n \cup \{\psi_n\} \text{ is S-consistent} \\ \Gamma_n \cup \{\Sigma_{\neg\psi_n}\} & \text{otherwise} \end{cases}$$

Now let $\Gamma = \bigcup_{n \geq 0} \Gamma_n$. Clearly, $\Gamma \vdash_S \delta$ for every $\delta \in \Delta$. Moreover, Γ_0 is S-consistent if Δ is, and the sequence $\Gamma_0, \Gamma_1, \Gamma_2, \ldots$ is defined so that if Γ_n is S-consistent, then Γ_{n+1} is S-consistent too. By induction on n, it follows that if Δ is S-consistent, then so is Γ_n for each n, and hence Γ itself.

To check that Γ is a Σ-maximally S-consistent set of Σ-instances, take $\psi \in \mathcal{L}$ and suppose $\Sigma_\psi \notin \Gamma$. ψ must appear somewhere in the list $\psi_0, \psi_1, \psi_2, \ldots$, that is, $\psi = \psi_m$ for some m. But if $\Sigma_{\psi_m} \notin \Gamma$, this must be because $\Gamma_m \cup \{\psi_m\}$ is S-inconsistent. And this means that $\Gamma_m \cup \{\Sigma_{\psi_m}\}$ is S-inconsistent, since

$$\vdash_S \psi_m \leftrightarrow \Sigma_{\psi_m}$$

Since $\Gamma_m \subseteq \Gamma$, $\Gamma \cup \{\Sigma_\psi\}$ is S-inconsistent too. □

 © Incurvati and Murzi 2017

Appendix B

For completeness, we show how McGee's result can be recovered as a special case of Theorem 5. In this case, we work with a countable first-order language $\mathcal{L}^\star$ in which we have relatively interpreted the language obtained by adding Tr to the language of arithmetic. We make use of the following lemma:

Lemma 9 (McGee). For each ϕ in $\mathcal{L}^\star$, there is a sentence B_ϕ such that

$$\vdash_{\mathsf{Q}} \phi \leftrightarrow (\mathsf{Tr}(\ulcorner B_\phi \urcorner) \leftrightarrow B_\phi)$$

Proof. Apply Tarski's Fixed-Point Theorem to the $\mathcal{L}^\star$-predicate $(\phi \leftrightarrow \mathsf{Tr}(x))$, and then use the associativity of biconditionals. □

Now let A be a consistent arithmetical theory which implies the axioms of Q. Then, in conjunction with Theorem 5, Lemma 9 immediately delivers McGee's result:

Corollary 10 (McGee). For any A-consistent set Δ of sentences in $\mathcal{L}^\star$, there is a (T)-maximally A-consistent set Γ of (T)-instances such that $\Gamma \vdash_{\mathsf{A}} \delta$ for every $\delta \in \Delta$. Moreover, for all $\psi \in \mathcal{L}^\star$, $\Gamma \vdash_{\mathsf{Q}} \psi$ or $\Gamma \vdash_{\mathsf{Q}} \neg\psi$.

References

Boolos, George 1971: 'The Iterative Conception of Set'. *Journal of Philosophy*, 68, pp. 215–14. Reprinted in Boolos 1998, pp. 13–29.

——1990: 'The Standard of Equality of Numbers'. In George Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, pp. 261–77. Cambridge: Cambridge University Press. Reprinted in Boolos 1998, pp. 202–219.

——1993: 'Whence the Contradiction?' *Proceedings of the Aristotelian Society Supplementary Volume* 67, pp. 213–33. Reprinted in Boolos 1998, pp. 220–36. Page references are to the reprint.

——1998: *Logic, Logic, and Logic*. Cambridge, MA: Harvard University Press.

Fine, Kit 2002: *The Limits of Abstraction*. Oxford: Oxford University Press.

Frege, Gottlob 1884: *Die Grundlagen der Arithmetik*. Breslau: W. Koebner.

 © Incurvati and Murzi 2017

Gödel, Kurt 1931: 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I'. *Monatshefte für Mathematik und Physik*, 38, pp. 173–98.

Goldstein, Laurence 2006: 'Fibonacci, Yablo, and the Cassationist Approach to Paradox'. *Mind*, 115, pp. 867–89.

——2012: 'Reasoning without Contradiction'. *The Reasoner*, 6, pp. 183–4.

——2013: 'Paradoxical Partners: Semantical Brides and Set-Theoretical Grooms'. *Analysis*, 73, pp. 33–7.

Grishin, Vyacheslav Nikolaevich 1981: 'Predicate and Set-Theoretic Calculi Based on Logic without Contractions'. *Mathematics of the USSR-Izvestiya*, 45, pp. 45–68.

Gupta, Anil, and Nuel Belnap 1993: *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Horwich, Paul 1990: *Truth*. 2nd edn. 1998. Oxford: Oxford University Press.

Incurvati, Luca 2014: 'The Graph Conception of Set'. *Journal of Philosophical Logic*, 43, pp. 181–208.

Leitgeb, Hannes 2005: 'What Truth Depends On'. *Journal of Philosophical Logic*, 34, pp. 155–92.

Maddy, Penelope 1988: 'Believing the Axioms I'. *Journal of Symbolic Logic*, 53, pp. 481–511.

McGee, Vann 1992: 'Maximal Consistent Sets of Instances of Tarski's Schema (T)'. *Journal of Philosophical Logic*, 21, pp. 235–41.

Quine, W. V. 1951: *Set Theory and Its Logic*. Cambridge, MA: Belknap Press.

Rosser, J. Barkeley 1936: 'Extensions of Some Theorems of Gödel and Church'. *Journal of Symbolic Logic*, 1, pp. 87–91.

Weber, Zach 2010: 'Transfinite Numbers in Paraconsistent Set Theory'. *Review of Symbolic Logic*, 3, pp. 71–92.

——2012: 'Transfinite Cardinals in Paraconsistent Set Theory'. *Review of Symbolic Logic*, 5, pp. 269–93.

Wittgenstein, Ludwig 1922: *Tractatus Logico-Philosophicus*. Translated by C. K. Ogden. London: Routledge and Kegan Paul.

Wright, Crispin 1983: *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

Zermelo, Ernst 1908: 'Untersuchungen über die Grundlagen der Mengenlehre I'. *Mathematische Annalen*, 65, pp. 261–81.

© Incurvati and Murzi 2017

What Verities May Be

Igor Douven
Sciences, Normes, Décision (CNRS), Paris-Sorbonne University
igor.douven@paris-sorbonne.fr

Lieven Decock
VU University Amsterdam
l.b.decock@vu.nl

Edgington has proposed a solution to the sorites paradox in terms of 'verities', which she defines as degrees of closeness to clear truth. Central to her solution is the assumption that verities are formally probabilities. She is silent on what verities might derive from and on why they should be probabilities. This paper places Edgington's solution in the framework of a spatial approach to conceptualization, arguing that verities may be conceived of as deriving from how our concepts relate to each other. Building on work by Kamp and Partee, this paper further shows how verities, thus conceived of, may plausibly be assumed to have probabilistic structure. The new interpretation of verities is argued to also help answer the question of what the verities of indicative conditionals are, a question which Edgington leaves open. Finally, the question of how to accommodate higher-order vagueness, given this interpretation, is addressed.

Natural languages are rife with vague predicates. Such predicates are generally taken to be characterized by their having borderline cases; by their supporting soritical reasoning; and by their having blurry boundaries, in that we cannot discern a sharp divide between their clear and their borderline instances.

The phenomenon of vagueness raises important philosophical questions. One is the question of how to account for the existence of borderline cases. Why is it that, for instance, some shades resist clear classification as red and equally resist clear classification as non-red? Another pressing question concerns the logic of vague languages. What is now commonly called 'classical logic' was devised as a tool for regimenting mathematical reasoning. However, the language of mathematics is special, in that no vagueness attaches to any of its predicates. So, a priori one might expect at least some friction if classical logic is applied to natural language. Many have taken the sorites paradox to show that, in effect, the problem runs much deeper, and that classical logic can be applied to natural language

Mind, Vol. 126 . 502 . April 2017 **© Douven and Decock 2017**
doi:10.1093/mind/fzv194 Advance Access publication 19 January 2017

only by sacrificing consistency. But then which logic does apply to natural language? Or might classical logic still apply, albeit with certain qualifications, perhaps?

Edgington (1997a) proposes a degree-theoretic account of vagueness by introducing 'verities', which are defined as degrees of closeness to clear truth, and which are supposed to have probabilistic structure. Edgington's account yields a highly elegant solution to the sorites paradox, which not only avoids various infelicities incurred by rival degree-theoretic approaches to vagueness, but also leaves a well-delineated, only slightly restricted, role for classical logic in reasoning outside of mathematics.

Edgington's proposal raises a number of issues of its own, however. As Edgington (1997a, p. 307) admits, she only *postulates*, and does not prove, that verities obey the axioms of probability. More generally, one may wonder what the metaphysical status of verities is; that they help solve the sorites paradox gives at best a very inconclusive reason for believing in their reality.[1] A further question concerns the verities of indicative conditionals, a question that Edgington flags, only to set it aside. The question is pressing — supposing, as Edgington also does, that the semantics of indicative conditionals is not given by the truth table for the material conditional — for there exist versions of the sorites paradox that have indicative conditionals among their premisses.

This paper aims to provide a foundation for Edgington's solution to the sorites paradox by giving an account of what verities may be. The account is plausible inasmuch as it provides satisfactory answers to the questions left open by Edgington, but also inasmuch as it connects with the scientific literature on human concept formation.[2]

The basic idea underlying our proposal is simple enough and is not original with us. Lewis (1970), Kamp (1975), Kamp and Partee (1995), and Akiba (2004) have already outlined how degrees of truth can be

[1] Fara (2000, p. 54) makes the general point that, absent a 'substantial philosophical account of what degrees of truth are', we have no reason to accept degree-theoretic approaches to vagueness.

[2] For an earlier attempt to provide a metaphysical grounding for Edgington's verities, see van Deemter (2010, pp. 210–13), which proposes to understand verities in terms of people's judgements of instances of concepts. What we find particularly attractive about our account is that it flows naturally from an independently motivated theory of concepts. Van Deemter's account does not appear to have a comparable virtue. Also, given our account, it is straightforward to *explain* the kind of judgements on which van Deemter's account relies. Van Deemter seems to take these judgements as being somehow basic; at any rate, he makes no attempt to explain them.

 © Douven and Decock 2017

defined in a measure-theoretic way on the basis of a supervaluationist approach to vagueness.[3] As is so often the case, however, the devil is in the details. Kamp and Partee, who have gone furthest in fleshing out the measure-theoretic idea, admit that, in the end, their proposal fails to deliver a unique measure of degrees of truth; they also admit to being clueless as to how to arrive at a unique measure.

Implementing Kamp and Partee's proposal in a spatial approach to conceptualization, and concerned with the membership relation of fuzzy set theory rather than with (graded) truth, Decock and Douven (2014) shows how to obtain, in a principled manner, a unique measure of graded membership from sets of supervaluations. As will be seen, this measure rather straightforwardly suggests a definition of graded truth for atomic sentences. The main technical tasks of the present paper are, first, to extend that definition to compound sentences, again in a principled manner, and second, to prove that, thus defined, degrees of truth have probabilistic structure, as required for Edgington's solution to the sorites paradox to work. At a metaphysical level, the resulting picture will be seen to illuminate the nature of verities; in this picture, verities flow from the structure of conceptual reality. The picture will also help answer the question about the verities of indicative conditionals.

The account of verities to be offered may at first make it appear harder to account for higher-order vagueness, which is a now common name for the fact that vague predicates have blurry boundaries.[4] For Edgington, the precision that comes with the probabilistic structure of verities, and with degree-theoretic approaches to vagueness more generally, is artefactual, to be considered as a feature of her model that does not correspond to any aspect of reality. That is her answer to the concern that her proposal fails to do justice to phenomena of higher-order vagueness. In the account to be offered, the

[3] In her 1992 paper, which in many ways anticipates the longer Edgington (1997a) on vagueness and the sorites, Edgington is at least sympathetic to the broad idea of defining verities along these lines. In the latter paper, she distances herself from the idea again, without explaining why. Keefe (2000, p. 171) also expresses sympathy for degree-theoretic supervaluationism, although she stops short of endorsing it, and does not draw on degrees of truth in her treatment of the sorites paradox. She may want to reconsider in view of Fara's (2010) critique that the brand of non-degree-theoretic supervaluationism that Keefe endorses fails to supply a satisfactory psychological explanation of why the premisses of the sorites paradox strike us as true. As will be seen, this is where degrees of truth can play a particularly useful role.

[4] For present concerns, we are granting the existence of cases of higher-vagueness. At the same time, we agree with Dietz (2011, pp. 142–3) that the evidence for higher-order vagueness is not nearly as strong as that for first-order vagueness.

 © Douven and Decock 2017

precision of verities is real, and differences in verities, however slight, reflect differences in reality. We argue that, nevertheless, our account has no difficulty accommodating higher-order vagueness.

We start by expounding Edgington's proposed solution to the sorites paradox, in §1. In §2, we describe an approach to conceptualization, developed in the cognitive sciences, that represents concepts spatially. In the same section, we describe earlier work of our own which relies on that approach and will provide the framework for defining verities in terms of conceptual structures. Then, in §3, we state the definition of verities and show that it entails verities to have probabilistic structure. We also there address the question of the verities of indicative conditionals and respond to the concern that the framework has difficulty dealing with higher-order vagueness.

1 Edgington on the sorites

Degree-theoretic approaches to vagueness all aim to handle vagueness by admitting as semantic values for propositions, next to truth (1) and falsity (0), continuum many intermediate degrees of truth, usually represented as real numbers in the interval (0, 1). Standard degree-theoretic approaches are degree-functional in that in these approaches, the degree of truth of a compound sentence is a function of the degrees of truth assigned to its constituent parts. The best known approaches of this kind have adopted Łukasiewicz's generalizations of the truth tables for the binary propositional operators, meaning, among other things, that they define the degree of truth of a sentence's negation to equal 1 minus the degree of truth of the original sentence; the degree of truth of a conjunction to equal the minimum of the degrees of truth of its conjuncts; and the degree of truth of a disjunction to equal the maximum of the degrees of truth of its disjuncts.

As various authors have pointed out, these definitions have some counter-intuitive consequences. For instance, it follows from them that contradictions can have non-zero degrees of truth. If Bob is borderline tall, then, we may suppose, 'Bob is tall' will have some non-extreme degree of truth. Hence, its negation will equally have a non-extreme degree of truth. But then so will the conjunction of these sentences, 'Bob is tall and it is not the case that Bob is tall', contrary to one's reasonable expectation that it, and contradictions in general, will turn out to have a degree of truth of 0. It similarly follows that tautologies can have degrees of truth of less than 1.

 © Douven and Decock 2017

For Edgington, these and similar infelicities supply grounds to abandon the idea of degree-functionality. As she notes, abandoning degree-functionality does not automatically render one unable to say anything systematic about the relation between the degrees of truth of compound sentences and the degrees of truth of the parts of those sentences. Compare probability theory: the probability of a conjunction is not generally a function of the probabilities of the conjuncts, due to the fact that there may be dependencies between the conjuncts; for the same reason, the probability of a disjunction is not generally a function of the probabilities of the disjuncts. Nevertheless, probabilities of conjunctions obey the strict principle that $\Pr(\phi \wedge \psi) = \Pr(\phi)\Pr(\psi|\phi)$, and probabilities of disjunctions the equally strict principle — which in most presentations of probability theory has the status of an axiom — that $\Pr(\phi \vee \psi) = \Pr(\phi) + \Pr(\psi) - \Pr(\phi \wedge \psi)$.

In Edgington's analysis, the verity of a conjunction is similarly not (in general) a function of the verities of the conjuncts, because the verity of the conjunction may also depend on how the conjuncts relate to one another. To take into account such possible dependencies when determining the verity of a conjunction, we must, on her proposal, also consider what verity a conjunct has, on the supposition that we decide to count the other conjunct as being clearly true, which gives us the *conditional verity* of the former conjunct given the latter. *Supposing* that we decide to count as clearly true that Bob is tall, then that Bob is *not* tall has a verity of 0.

Edgington postulates a parallel between probabilities and verities that runs deeper than this: according to her, we should think of verities as having probabilistic structure generally. This leaves her in agreement with earlier proposals on the clause for negation but not with those for conjunction and disjunction. Letting Ver denote the verity operator and using $\mathrm{Ver}(\psi|\phi)$ as notation for the conditional verity of ψ given ϕ, in Edgington's proposal we have $\mathrm{Ver}(\phi \wedge \psi) = \mathrm{Ver}(\phi)\mathrm{Ver}(\psi|\phi)$ and $\mathrm{Ver}(\phi \vee \psi) = \mathrm{Ver}(\phi) + \mathrm{Ver}(\psi) - \mathrm{Ver}(\phi \wedge \psi)$. As Edgington (1997a, p. 306) notes, these clauses are also generalizations of the standard truth tables for conjunction and disjunction.

To be more exact, on Edgington's proposal, probability theory — which had already been recruited for axiomatizing chances, relative frequencies, and rational credences (degrees of belief) — can equally serve to axiomatize verities, for instance, in the following form:

$$(1)\quad 0 = \mathrm{Ver}(\phi \wedge \bar{\phi}) \leq \mathrm{Ver}(\phi) \leq \mathrm{Ver}(\phi \vee \bar{\phi}) = 1$$

© Douven and Decock 2017

$$(2)\quad \mathrm{Ver}(\phi \vee \psi) = \mathrm{Ver}(\phi) + \mathrm{Ver}(\psi) - \mathrm{Ver}(\phi \wedge \psi)$$

Following standard practice in probability theory, we can also define $\mathrm{Ver}(\psi|\phi)$ to equal $\mathrm{Ver}(\phi \wedge \psi)/\mathrm{Ver}(\phi)$.[5]

Classical logic determines which propositions are tautologies—these get a verity of 1—and which are contradictions—these get a verity of 0. Above and beyond this, does Edgington's account leave any role for classical logic? It does. Adams (1975) defines a notion of validity in terms of high probability preservation. An argument is valid, in his sense, iff the improbability of the conclusion (i.e. 1 minus the probability of the conclusion) cannot exceed the sum of the improbabilities of the premisses; or, to put it a bit more informally (and less precisely), iff the conclusion is highly probable whenever the premisses are highly probable. Adams further proves that an argument is classically valid iff it is valid in his sense. Edgington (1997a, p. 302) defines a parallel notion of validity in terms of high verity preservation. An argument is valid, in her sense, iff the unverity of the conclusion (1 minus the verity of the conclusion) cannot exceed the sum of the unverities of the premisses; more informally, iff the conclusion is close to being clearly true whenever the premisses are close to being clearly true. Edgington (1997a, §4) proves that all classically valid arguments are valid in her sense, too, and vice versa.

Of course, expanding our set of beliefs by means of classical reasoning is fine when the reasoning involves only certainties as premisses. In view of Adams's result, however, it is also fine when the premisses are only near certainties, except that then there is a proviso, to wit, that there not be too many premisses. Kyburg's (1961) lottery paradox brings out the need for the proviso. Each application of the rule of conjunction introduction may bestow a bit more uncertainty on the conclusion, and so when the rule is applied over and over

[5] At various junctures in her work on conditionals, Edgington has stressed that the standard definition of conditional probability should not be taken to suggest that credences of conjunctions are conceptually prior to conditional credences. According to her, it is often the other way around: we first determine the conditional credence and then, with its help, determine the credence to be attributed to the conjunction. *How* we determine conditional credences in such cases is, according to Edgington, largely captured by a footnote in Ramsey (1929), which suggests that we fix the conditional credence of ψ given ϕ by hypothetically adding ϕ to our stock of beliefs and then judging, from within that hypothetical belief state, what credence to assign to ψ (see, for instance, Edgington 1995, pp. 266–7; 1997b, pp. 108–9). Edgington may think that something similar holds for conditional verities. At any rate, in Edgington (1997a) she does not formally define conditional verity via the ratio formula, but only gives the informal characterization mentioned in the text above. (In Edgington 1992 she does define conditional verity via the ratio definition.)

 © Douven and Decock 2017

again, one may end up with a conclusion that is highly uncertain, or even certain to be false — as in the lottery paradox the conclusion that no ticket will win, where the lottery is known to have a winner — even though one's premisses were all near certainties — as in the lottery paradox the premisses to the effect that a given ticket will lose. Importantly, conjunction introduction is a valid rule of inference, classically and hence also in Adams's sense; it is the unconstrained use of the rule in a context in which we deal with near certainties, and not (only) perfect certainties, that can spell trouble.

Similarly, classical reasoning is fine, not only when applied to premisses that have a verity of 1, but also when applied to premisses that have a verity that is only close to 1, again given the proviso that there not be too many premisses. Here, the sorites paradox may serve to bring out the need for the proviso. Imagine a horizontally ordered series of n colour patches, with the leftmost patch being clearly red, and with patches almost unnoticeably becoming more orangish as we progress to the right, until we reach the rightmost patch, which is clearly orange. Let these patches be numbered 1 through n, from left to right. Now consider this argument involving these patches:

P_0 Patch number 1 is red.
P_1 If patch number 1 is red, then so is patch number 2.
P_2 If patch number 2 is red, then so is patch number 3.
⋮ ⋮
P_{n-1} If patch number n - 1 is red, then so is patch number n.
∴ Patch number n is red.

P_0 is undeniable. P_1 through P_{n-1} may not be equally undeniable, but pre-theoretically they appear true as well, in view of the fact that adjacent patches differ almost imperceptibly in colour. Yet the conclusion, which is reached by $n - 1$ applications of *modus ponens*, is a manifest falsehood. In Edgington's analysis of what has gone wrong, $\mathrm{Ver}(P_0) = 1$, as it should. She further proves that, supposing P_1 through P_{n-1} to be material conditionals, $\mathrm{Ver}(P_i) \approx 1$ for all $i \in \{1, \ldots, n-1\}$. Also, *modus ponens* is valid in her sense in that it preserves the property of being close to clearly true. That we reach a false conclusion by applying a valid rule to a series of premisses all of which have at least a verity close to 1 is because we are flouting the proviso: the reasoning relies on too many premisses which are only close to being clearly true, and the many slight unverities of these

 © Douven and Decock 2017

premisses accumulate in the conclusion, even to such an extent as to make it clearly false.[6]

Edgington's solution to the sorites paradox is very elegant, making maximally good sense of our pre-theoretical judgements about the truth of the premisses — given that they all have a verity at least close to 1, it is not surprising that they strike us as being true[7] — the falsity of the conclusion, and the validity of *modus ponens* and classical reasoning more generally. This, together with the fact that Edgington is able to avoid the infelicities that beset other degree-theoretic accounts, should be enough to recommend Edgington's proposal to anyone who is attracted to the idea that vagueness is to be understood by reference to degrees of truth.

But why should anyone want to concede that this is the right approach to begin with? Given Edgington's proposal, there are many parallels between credences and verities, but credences and verities are also asymmetrical in important respects. Most notably, credences are generally taken to have a straightforward operationalization in terms of betting odds. For all Edgington shows, verities do not have any operational definition at all. That need not be a problem. While behaviourists would have protested, nowadays we have no difficulty acknowledging that introspection offers strong evidence for our believing different propositions to different degrees. Verities might have the same phenomenological plausibility. In our opinion, this is simply not so, and the widespread resistance that degree-theoretic approaches to vagueness have encountered suggests that our opinion

[6] Edgington (1997a, pp. 311–12) also considers the shorter version of the sorites paradox that replaces $P_1, \ldots, P_{n-1}$ by the single premiss 'For all $i \in \{1, \ldots, n-1\}$, if patch number i is red, then so is patch number $n+1$'. She argues that the long version is to be seen as the basic form, and that the universally quantified premiss of the short version is to be understood by reference to $P_1, \ldots, P_{n-1}$, which serve as its grounds. According to Edgington, these grounds each bestow some slight amount of unverity on the universally quantified premiss, but jointly they bestow enough unverity on it to make it clearly false. So, in the short version too, the paradox embodies valid reasoning; it is just that it has a false premiss. That this premiss still strikes us as true is because each of its instances has a verity very close to 1.

[7] Note that this is to make a psychological claim, about how things appear to us, and not a normative one; in particular, we are not claiming that a verity close to 1 is sufficient for truth. Some passages in Edgington (1997a) seem to indicate that she is sympathetic to taking that further step and claiming that the premisses are true *tout court*. For she suggests that verities stand to plain truth as (according to many) credences stand to the notion of categorical belief, and that believing a given proposition to a degree of 1 is sufficient, but not necessary, for qualifying as categorically believing that proposition. Edgington is not very explicit about plain truth, however, and it is possible to explain why the premisses of the sorites paradox appear true to us without making assumptions that, perhaps, she would not countenance after all.

 © Douven and Decock 2017

on this matter is not particularly idiosyncratic. Verities could still be explanatorily indispensable, and could be justified for that reason. However, that is hard to maintain in the presence of solutions to the sorites paradox that do without verities — even if none of these may be as elegant as Edgington's — and in the absence of any indication of what further work verities might be able to do for us. Thus verities appear to rest on a somewhat tenuous footing.

Even granting that verities exist, there is still an asymmetry between them and credences, in that there are well-established arguments — Dutch book arguments and accuracy minimization arguments — to the effect that our degrees of belief ought to satisfy the probability axioms, on pain of irrationality. No argument is known to support the contention that verities ought to satisfy the probability axioms.

There is a further concern that one may have about Edgington's solution. As mentioned, Edgington interprets P_1-P_{n-1} as material conditionals. She is surely right to assert that it is important to have a solution to versions of the paradox in which the conditionals are material conditionals (Edgington 1997a, p. 307), but one would also like to know how to deal with versions of the paradox that proceed on the basis of indicative conditionals, which nowadays are almost generally believed not to share their semantics with material conditionals.[8] It is not immediately obvious that Edgington's solution should work for such versions as well. After all, if indicative conditionals are stronger than their material counterparts — as most theorists hold[9] — then they may well have lower verities in general than those counterparts. So it could be that the verities of the conditionals in the sorites paradox, when read as indicative conditionals, are *not* close to 1 — which again would make it puzzling why they at least *appear* to be true. Whether there is any ground for this concern is a question we will only be able to answer if we know more about the verities of indicative conditionals.

Edgington (1997a, p. 307) makes the intriguing remark that it is 'tempting' to postulate that, for all propositions ϕ and ψ,

$$\text{(EQ)} \quad \mathrm{Ver}(\text{If } \phi, \text{ then } \psi) \quad = \quad \mathrm{Ver}(\psi|\phi)$$

[8] Edgington herself is one of the staunchest critics of the view that indicative conditionals have the same truth conditions as their material counterparts; see, for example, Edgington (1995).

[9] Lycan (2001, ch. 3) and Krzyżanowska, Wenmackers and Douven (2013, 2014) are notable exceptions.

 © Douven and Decock 2017

where 'If ϕ, then ψ' is an indicative conditional. The remark appears in brackets, and she does not expand on it. Why only tempting? Recall that $\mathrm{Ver}(\psi|\phi)$ stands for the verity ψ has on the supposition that ϕ is counted as clearly true. In so far as the if-clause of the corresponding conditional expresses that supposition, (EQ) seems to be backed by intuition (if not, perhaps, by formal argument).

Edgington's (1997a) apparent hesitation regarding (EQ) is the more remarkable given that in Edgington (1992, p. 208) she does seem to endorse the principle, albeit more or less in passing. On the other hand, it is not so hard to guess what will have given her second thoughts in between writing the earlier paper and writing the later one. If, as Edgington argues, the verity operator, Ver, and the probability operator, Pr, have the same formal features, then anyone familiar with the literature on indicative conditionals will fear that Lewis's and others' triviality arguments against the thesis — often called 'Stalnaker's Hypothesis' — that the probabilities of indicative conditionals equal the corresponding conditional probabilities, carry over to (EQ). If these arguments do carry over, then (EQ) has some manifestly false consequences, such as that, for all ϕ and ψ such that $\mathrm{Ver}(\phi \wedge \psi) > 0$ and $\mathrm{Ver}(\phi \wedge \bar{\psi}) > 0$, $\mathrm{Ver}(\psi|\phi) = \mathrm{Ver}(\psi)$; that is, barring special cases, conditional verities equal unconditional ones.[10] This would certainly be a good reason not to commit oneself to (EQ).

The triviality arguments for the probabilities of conditionals require assumptions that go beyond the analogue of (EQ). For instance, Lewis's (1976) triviality arguments assume that that analogue holds for a class of probability functions, where this class is to contain at least two different probability functions. The same holds for most of the other triviality arguments to be found in the literature on conditionals. The triviality arguments for which it does not hold make other additional assumptions, like the possibility of representing people's belief states by means of finitely many possible worlds (as in Hájek's 1989 argument), or the possibility of particular operations on probability spaces (like moving around probability mass, as in Hájek's 1994 argument). Whether these additional assumptions make sense for verities is difficult to assess so long as we are in the dark about what verities are. If they do not make sense, then, for all anyone has shown, we are free to take (EQ) on board. And then, given that, as

[10] We are granting here that the triviality arguments against Stalnaker's Hypothesis are cogent. For some reasons to doubt that they are, see Douven (2013; 2016a, ch. 3), Douven and Verbrugge (2013).

 © Douven and Decock 2017

Edgington shows, Ver(patch number $i + 1$ is red| patch number i is red) is close to 1 for all patches $i \in \{1, \ldots, n - 1\}$ in the soritical colour series considered above, her solution will also hold for the version of the sorites paradox with indicative conditionals.

Finally, Edgington's solution to the sorites paradox has raised the concern that, for all the aforementioned virtues, it runs aground on the problem of higher-order vagueness. The concern has been voiced with respect to degree-theoretic approaches to vagueness generally. At the core of the concern is the observation that, on all such accounts, there are, for all vague predicates F, *sharp* divides between clear cases of F-ness — those objects a such that F(a) gets a degree of truth (or degree of closeness to clear truth, as on Edgington's account) of 1 — clear non-cases of F-ness — those objects a such that F(a) gets a degree of truth of 0 — and borderline cases of F-ness, being objects a such that F(a) gets an intermediate degree of truth. In reality, however, the question of whether something is a borderline instance or a clear (non-)instance of a given predicate often lacks a clear answer itself; vagueness can impinge on judgements of vagueness, which gives rise to phenomena of higher-order vagueness.

Edgington is aware of these phenomena, of course, but she deals rather cursorily with them. Her response boils down to the claim that the model of vagueness she outlines is precisely that: a model. As such, it helps us to get a grip on vagueness even if some aspects of it are unrealistic or only vaguely related to reality.[11] She writes:

> The relation between the representation and the reality is still vague: there need be no fact of the matter exactly what number to assign. Worthwhile results generated by the idealisation must be robust enough to be independent of small differences. (Edgington 1997a, p. 297)

Probability theory provides a useful model of rational people's belief states even if in reality these belief states are much cruder than the model makes them seem. Very small differences in credences assigned to propositions may be just noise, and may not correspond to anything in people's minds or their dispositions to act.

Cook (2002), who is favourably inclined toward Edgington's degree-theoretic solution to the sorites paradox, provides on her behalf a more elaborate and refined response to the higher-order vagueness challenge. He does so by presenting a general picture of

[11] See also Williamson (1994, p. 131) for the suggestion that the degree theorist ought to respond to higher-order vagueness by claiming that the assignment of degrees of truth to propositions is itself subject to vagueness.

 © Douven and Decock 2017

semantic modelling in which semantic models are not meant to describe each and every detail of a given part of linguistic reality, but are rather meant to serve as 'one tool among many that can further our understanding of the discourse in question' (Cook 2002, p. 234). In this picture, semantic models contain elements that are intended to represent features of reality ('representors', as Cook calls them) as well as elements that are just there for mathematical convenience (called 'artefacts'). Specifically with respect to Edgington's model of vagueness, Cook's claim is that while verities are representors, the ultra-fine distinctions between verities that the model countenances are not.

Keefe (2012) has recently criticized Cook's distinction between representors and artefacts, arguing that there is no principled basis for the distinction.[12] Be this as it may, we shall offer a different response to the higher-order vagueness problem on Edgington's behalf. This response makes, her own suggestions to the contrary notwithstanding, the precision inherent in Edgington's model fully representative: each difference in verities, even if vanishingly small, corresponds to a difference in conceptual reality (in a sense that will be made clear below). Higher-order vagueness is then explained by reference to our limited capacities to survey the relevant aspects of that reality.

2 From conceptual spaces to graded membership

When introducing and motivating verities, Edgington approvingly cites Sainsbury's (1991, p. 182) suggestion that we may think of concepts as 'like magnetic poles exerting various degrees of influence'. She illustrates this suggestion by saying 'There are clearly red things, and clearly orange things', and 'In between, there are things which are closer to or further from the red pole'. Instead of thinking that there is a 'cut-off point' where objects cease to fall under a concept, 'we should think in terms of varying degrees of closeness to clear cases' (Edgington 1997a, p. 296).

In our view, both in thinking of concepts in spatial terms — 'in between', 'closer to', 'further from' — and in seeing the idea of verities as related to this spatial conception of concepts, Edgington is exactly right. What she may not have realized is that, at the time of her writing, a spatial conception of concepts had already been developed in the cognitive sciences, for completely independent reasons. In the

[12] See Smith (2012) for further criticisms of Cook's proposal.

 © Douven and Decock 2017

next section, we show that this conception of concepts provides a foundation for Edgington's proposal from which the probabilistic structure of verities that Edgington postulates flows naturally. The present section describes in some detail the spatial conception of concepts and relevant related work, in particular, work showing how to account for the existence of borderline cases given this conception, and a theory of graded membership that capitalizes on that work.

2.1 The conceptual spaces approach

Over the past thirty or so years, in psychology and in cognitive science more generally, a spatial view of concepts has emerged. According to this view, concepts are to be thought of *more geometrico*, as regions in conceptual spaces. For example, in this view the concept RED is not to be identified with a set of necessary and sufficient conditions for qualifying as red, nor with a set of conditions a large enough subset of which must be satisfied for something to qualify as red, nor with abilities of agents to discriminate red things from non-red things, nor with anything else that has hitherto been proposed in the traditional philosophical literature. Instead, RED is a region of colour space, where this is generally conceived as a three-dimensional metrical space with one dimension representing hue, one dimension representing brightness, and one dimension representing saturation. Similarly, SALTY is a region of (three-dimensional) taste space, and HIGH (as pertaining to sounds) a region of (two-dimensional) auditory space.

More generally, in this account concepts are regions of one- or multidimensional metrical spaces, now usually called 'conceptual spaces'. Mathematically speaking, a metrical space is any set of points with a distance function (or metric) defined on it, where $\delta : S \times S \rightarrow \mathbb{R}$ is said to be a distance function on space S iff, for all points $a, b, c \in S$, (i) $\delta(a, b) \geq 0$, with $\delta(a, b) = 0$ iff $a = b$; (ii) $\delta(a, b) = \delta(b, a)$; and (iii) $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$. In conceptual spaces, the dimensions are taken to represent fundamental qualities in terms of which objects may be compared to each other. Objects are mapped onto points in conceptual spaces according to the values they assume with respect to the spaces' various quality dimensions.

In this picture, distances between objects in a given space — or rather, the distances between the points representing the objects in the space[13] — are interpreted as measuring the dissimilarity of the

[13] Typically, we are not meticulous about distinguishing between an object, the point in a space (or the points in spaces) onto which the object is mapped, and the name of the object.

 © Douven and Decock 2017

objects. To be more exact, the similarity of two objects in a given respect is supposed to be a monotonically decreasing function of the distance metric defined on the space associated with that respect: the closer their representations in the space, the more similar the objects are in the corresponding respect. For example, two objects that are close in colour space are more similar in colour than two objects that are far apart in the space.

In the literature, it is commonly assumed that the Euclidean metric familiar from high-school geometry provides the appropriate metric for spaces whose dimensions are *integral*, meaning that their dimensions correspond to qualities that cannot be perceived separately. (For instance, it is not possible to hear the pitch of a tone without hearing its loudness.) It has been argued, however, that the city block metric, which basically just adds up distances along all dimensions, is a better choice of metric for spaces with *separable* (i.e. non-integral) dimensions.

We have already mentioned colour space, taste space and auditory space as examples of conceptual spaces. Many other conceptual spaces have been described in the literature, including olfactory space, various shape spaces, actions spaces, and value spaces. In fact, the arsenal of conceptual spaces continues to grow. For instance, Gärdenfors and Zenker (2011, 2013) have recently developed a number of conceptual spaces for physical concepts such as MASS and FORCE, and a prominent example in Churchland (2012), which investigates the possible neural basis of conceptual spaces, is that of a high-dimensional face space (for the classification of human faces). How far this approach to conceptualization can be taken, and whether all concepts can eventually be brought under its umbrella, remains to be seen. For present purposes, the important point to notice is that the conceptual spaces approach has so far been applied with great success to phenomenal concepts, which are the concepts most closely associated with problems of vagueness and are typically encountered in versions of the sorites paradox.

To forestall misunderstanding, we should note that it would be wrong to think of conceptual spaces as being mere devices for visually representing concepts, which may facilitate theorizing and communicating about concepts, but whose geometrical or topological structures are otherwise arbitrary. On the contrary, these spaces are constructed on the basis of empirical data by means of largely uncontentious statistical techniques, most notably a set of techniques known as 'multidimensional scaling'. Often the data that underlie the

 © Douven and Decock 2017

construction are similarity judgements, as for instance when people have been asked to judge the similarity of one colour patch to another, for sufficiently many pairs of patches. Conceptual spaces can also be based on less direct measures, as when frequencies have been registered at which people mistake one colour patch for another, or how long it takes them to discern a difference between specific colour patches. Such data are used to construct a so-called similarity matrix, which serves as the input for the designated statistical techniques.[14]

As stated previously, in the present view, concepts are regions of conceptual spaces, where a region of a space is a subset of points of that space. While, in principle, any region could represent a concept, only a relatively small number of them correspond to concepts that actually figure in our thinking. Gärdenfors (2000, pp. 70–7) shows that the concepts we use—'natural concepts', as he calls them—tend to have a special topological feature, to wit, they are *convex* regions in conceptual spaces, which is to say that for every pair of points in them, the line segment connecting the points lies in its entirety within the concept.

Gärdenfors (2000, pp. 87–91) makes a strong case for combining the framework of conceptual spaces with prototype theory, which is (roughly) the view that among the instances of a concept, some are more representative of it than others, with the most representative instance being the *prototype* of the concept. According to Gärdenfors, this combination yields a highly plausible account of categorization, at least when further conjoined with the mathematical technique of Voronoi diagrams (also known as Voronoi tesselations). A Voronoi diagram on a given space divides that space into cells according to the principle that each cell has a 'centre', and further contains all and only those points that lie no closer to the centre of any other cell than to its own centre. Formally, with S an m-dimensional metrical space with metric δ_S and $P = \{p_1, \ldots, p_n\}$ a set of pairwise distinct points in S, the region

$$v(p_i) := \left\{p \mid \delta_S(p, p_i) \leq \delta_S(p, p_j), \text{ for all } j \in \{1, \ldots, n\} \text{ with } j \neq i\right\}$$

[14] For more on the construction of conceptual spaces, we refer the interested reader to the highly informative appendix of Clark (1993). Borg, Groenen and Mair (2013) is a more technical, yet still very accessible, introduction to the statistical techniques involved in the construction of conceptual spaces

 © Douven and Decock 2017

is the *Voronoi polygon or polyhedron associated with* p_i, and the set

$$V(P) := \{v(p_i) | p_i \in P\}$$

is the *Voronoi diagram generated by P*. Fig. 1 shows a Voronoi diagram on a two-dimensional Euclidean space. It is easy to see that all Voronoi polygons in this diagram are convex. This is not peculiar to this diagram: Okabe et al. (2000, p. 58) prove that all Voronoi polygons and polyhedrons of Voronoi diagrams on Euclidean spaces are convex. Thus, at least for the most common type of conceptual space, the Voronoi diagram on the space generated by the prototypical points will automatically carve up that space into regions which share at least one important property (convexity) with natural concepts. Gärdenfors (2000, 2014) musters a welter of empirical evidence suggesting that the described procedure is in fact more generally along the right lines, and that it tends to produce regions largely coinciding with concepts in use (see also Douven 2016b and Douven *et al.* 2017).

2.2 Borderline vagueness

At the outset, it was mentioned that one of the central questions in the debate about vagueness is the question of what a borderline case is. At first blush, the account of conceptualization we are considering suggests a perfectly straightforward answer to this question: a borderline case of a given concept is a case that lies equally as far from the prototype of the concept as it does from the prototype of at least one other concept represented in the same space. On more careful consideration, however, this answer cannot be entirely right. After all, there can be, say, borderline cases of YELLOW/ORANGE such that *any* minute change in their colour would again result in a YELLOW/ORANGE borderline case. But this could not hold true if borderlines were really just one point 'thick', as they are in the present account. 'Almost all' small changes in colour—as small as one likes—of a YELLOW/ORANGE borderline case will result in either a clear case of yellow or a clear case of orange, given that all borderline cases are immediately adjacent to non-borderline cases.

This 'thickness problem' led Douven et al. (2013) to propose an extension of the conceptual spaces approach that yields conceptual spaces carved up by diagrams with thick borderlines. The proposal begins by observing that concepts may have non-unique prototypes. For instance, as Berlin and Kay's (1969) celebrated colour-naming experiments showed, people tend to designate more than one shade

 © Douven and Decock 2017

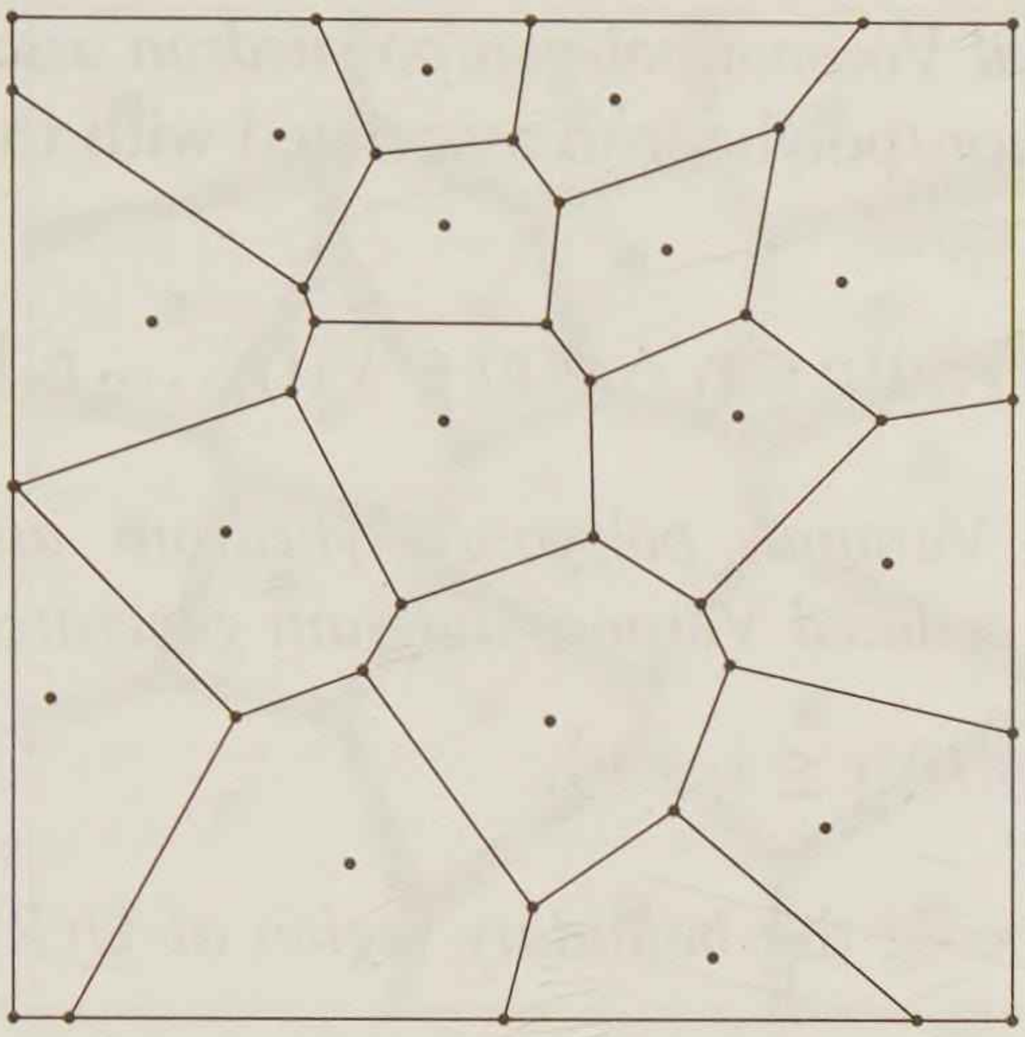

Figure 1: A two-dimensional Voronoi diagram

of any given colour when asked which shade is typical for the colour. In the conceptual spaces approach, this translates to the claim that conceptual spaces may have prototypical *regions* instead of prototypical *points*. As a next step, Douven et al. extend the technique of Voronoi diagrams to take into account the possibility of prototypical regions. In outline, the idea is to consider, for a given conceptual space S with prototypical regions $r_1, \ldots, r_n$, all n-tuples of points that pick precisely one point from each r_i. Each such n-tuple defines a Voronoi diagram on S. If one takes the set of all those diagrams and projects them onto each other, one gets (what Douven et al. call) the *collated Voronoi diagram on S*.

For later purposes, it is important to be precise about this. Let $R = \langle r_1, \ldots, r_n \rangle$ be an ordered set of prototypical areas in a given space S. Then

$$\Pi(R) \;:=\; \prod_{i=1}^{n} r_i \;=\; \left\{ \langle p_1, \ldots, p_n \rangle \,|\, p_i \in r_i, \text{ for } i \in \{1, \ldots, n\} \right\}$$

is the set of all ordered sequences $\vec{p} = \langle p_1, \ldots, p_n \rangle$ such that $p_i \in r_i \in R$ for $1 \leq i \leq n$, and

$$\mathcal{V}(R) \;:=\; \left\{ V(\vec{p}) \,|\, \vec{p} \in \Pi(R) \right\}$$

is the set of all Voronoi diagrams on S generated by elements of $\Pi(R)$. Letting

$$\underline{v}(p_i) \;:=\; \left\{ p \,|\, \delta_S(p, p_i) < \delta_S(p, p_j), \text{ for all } j \in \{1, \ldots, n\} \text{ with } j \neq i \right\}$$

© Douven and Decock 2017

define the *restricted Voronoi polygon/polyhedron associated with* p_i (i.e. the Voronoi polygon/polyhedron associated with that point minus its edges/faces), the region

$$\underline{u}(r_i) := \cap\{\underline{v}(p) | p \in r_i \wedge v(p) \in V(\{p_1, \ldots, p_n\}) \in \mathcal{V}(R)\}$$

is the *collated Voronoi polygon/polyhedron associated with* r_i. Furthermore, the *collated Voronoi diagram generated by* R is the set

$$\underline{U}(R) := \{\underline{u}(r_i) | 1 \leq i \leq n\},$$

with $S \setminus \bigcup \underline{U}(R)$ being the *boundary region* of $\underline{U}(R)$. And finally, the region

$$\bar{u}(r_i) := \bigcup_{p \in r_i} v(p)$$

is the *expanded collated Voronoi polygon/polyhedron associated with* r_i, and $\bar{u}(r_i) \setminus \underline{u}(r_i)$ the *boundary region associated with* r_i.

In the resulting version of the conceptual spaces approach, concepts can be identified with expanded collated Voronoi polygons/polyhedrons of conceptual spaces, which are composed of a restricted collated Voronoi polygon/polyhedron part and a boundary region part. Cases clearly falling under a concept are ones mapped onto points lying in the restricted part of the concept; borderline cases are ones mapped onto points lying in the boundary part.

Douven et al. prove that collated Voronoi diagrams have a number of mathematical properties which seem desirable from the perspective of providing an account of concepts. For instance, they show that concepts still turn out to be convex. Also, and this is crucial in relation to the thickness problem, borderline cases may be surrounded solely by borderline cases. Both features are clearly visible in the collated Voronoi diagram of a two-dimensional Euclidean space in Fig. 2.

2.3 Graded membership

Various authors have suggested modelling the phenomena of vagueness by means of fuzzy set theory, a theory developed by Zadeh (1965), Hájek (1998), and others, at the core of which is a notion of graded membership. In so far as designating a shade as borderline red is plausibly rendered as saying that the shade is a member of the class of red things only to *some* extent, the suggestion appears natural enough. If it has nevertheless found little acclaim, that is because

 © Douven and Decock 2017

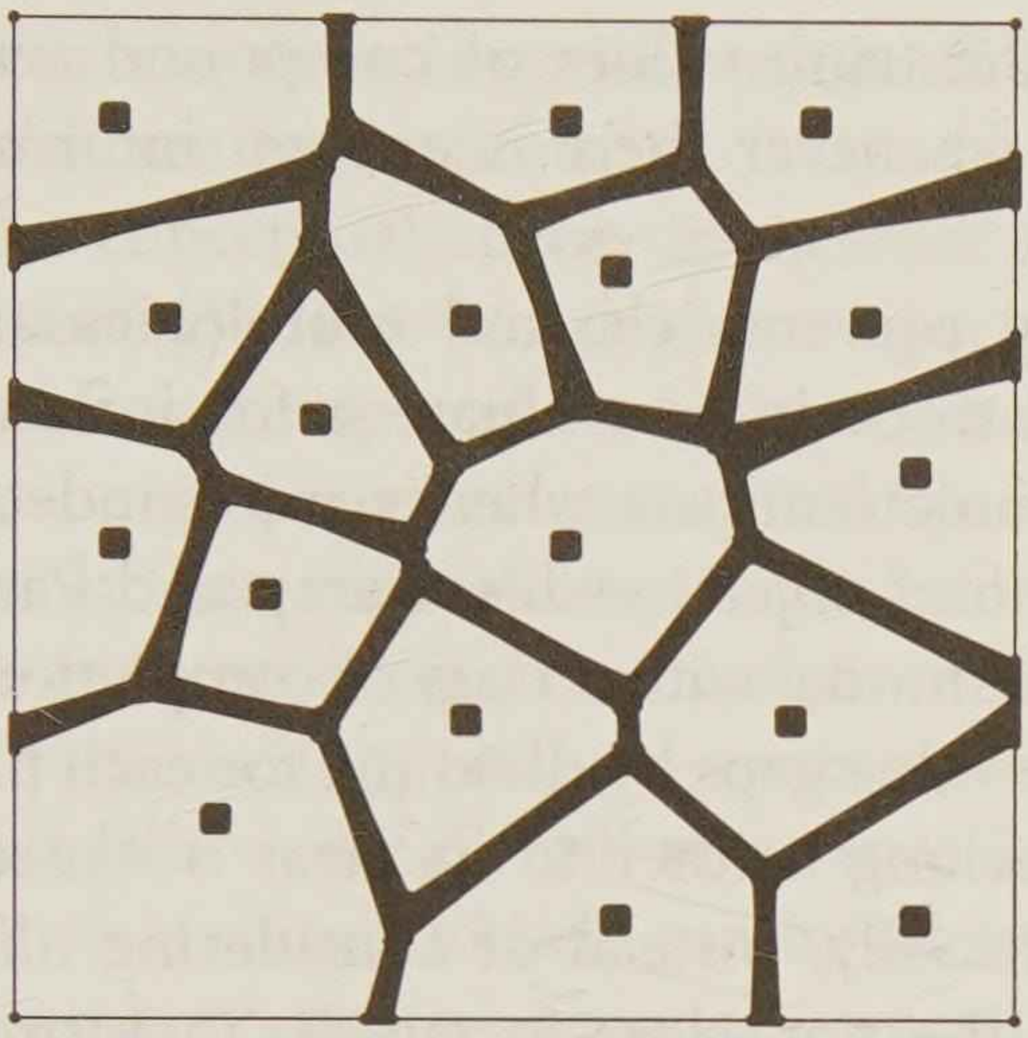

Figure 2: A two-dimensional collated Voronoi diagram

philosophers are generally wary of the notion of graded membership that is involved. As long as that notion is no better understood than the notion of a borderline case, what could the upshot of the formal exercise be?

Decock and Douven (2014) recruit the conceptual spaces framework, in the version summarized above, in service of developing an account of graded membership that is not only precise but also illuminating. They also draw on important work by Kamp and Partee (1995) which aims to provide a semantics for a language containing vague predicates.

More accurately, Kamp and Partee (1995, p. 148) assume a language with 'simple predicates' — roughly, predicates that have monolexemic expressions in English — which may be vague. They define a binary partial model for this language to consist of a universe of discourse (a set of individuals) and an interpretation function which, for each predicate in the language, divides the universe of discourse into a part that contains the clear instances of the predicate (the positive extension), a part that contains the clear non-instances of the predicate (the negative extension), and a part that contains the remaining objects (if this part is empty, the predicate is crisp).

Kamp and Partee (1995, pp. 148 ff.) define a partial semantics for their language in a number of steps. First, they assign to an atomic sentence F(a) the value 1 if a is a clear instance of F; 0 if a is a clear non-instance of F; and no value ('indeterminate' or 'undefined') if a is among the remaining objects. Then the strong Kleene scheme (Kleene 1952, §64) is invoked to extend the semantics to compound sentences;

 © Douven and Decock 2017

this means that the truth values of compound sentences are determined classically whenever there is enough information available to do so.

So far, nothing prevents classical tautologies and contradictions from turning out to be indeterminate — for instance, on the Kleene scheme, $\phi \vee \bar{\phi}$ is indeterminate whenever ϕ is indeterminate — which Kamp and Partee find objectionable. Kamp and Partee therefore supplement the partial model with a class of completions, which are ways to eliminate truth-value gaps by dividing, for each predicate, the set of individuals that belong neither to its clear instances nor to its clear non-instances. Crucially, instead of considering *all* logically possible ways of dividing the non-clear instances, they propose to consider only the ones — termed 'admissible' — which respect typicality rankings, in the sense that if they group a non-clear instance *a* of F with the clear F-instances, then they group anything with the clear F-instances that is at least as similar to the F-prototype as *a* is.[15] Kamp and Partee then use a supervaluationist approach to further extend the partial semantics. In this approach, a sentence receives the value 1 if it receives that value in all completions; it receives the value 0 if it receives that value in all completions; and its value is undefined otherwise. This ensures that all classical tautologies turn out to have the value 1 and all classical contradictions to have the value 0.

Finally, Kamp and Partee (1995, §4.3) suggest that, by setting the degree to which an object *a* falls under a given concept equal to the measure of the set of completions that group *a* with the clear instances of the concept, we can even assign something like intermediate truth-values to the sentences that so far still turn out to be indeterminate.[16] For instance, if the set of completions that group *a* with the F-prototype has measure 2/3, then that is also the truth-value of the sentence ⌜F(*a*)⌝.

This suggestion is made rather tentatively, though, for as Kamp and Partee note, the constraint to respect typicality orderings, even in conjunction with arguable formal constraints, is too weak to guarantee a unique measure on the set of completions, and thus too weak to guarantee a unique assignment of intermediate truth-values. As far as

[15] It is worth noting that Kamp and Partee take similarity as a primitive notion. In particular, they are not relying on the conceptual spaces account or any other formal account of similarity.

[16] The suggestion had been made before in the literature, as we saw in the introduction.

 © Douven and Decock 2017

Kamp and Partee can see, there is no way to obtain a unique measure without making arbitrary stipulations.

This is where the conceptual spaces framework may come to the rescue, as Decock and Douven (2014) show. Specifically, they show that through embedding Kamp and Partee's proposal in a conceptual spaces framework, the geometry of conceptual spaces overcomes the need to make any further stipulations to yield for each concept *C* representable in a conceptual space, a unique function — a graded membership function, as they call it — measuring the degree to which a given object falls under *C*. Still more specifically, the conceptual spaces framework is shown to help out by allowing the key elements of Kamp and Partee's proposal — partial models and completions — to be construed as geometrical objects which can themselves be represented in a space of the right dimensionality. The unique membership function is then obtained by metricizing that space in the intuitively and formally obvious way. Because the main claim of this paper (to be argued in the next section) depends on the details of this construction, it is well worth being more specific here.

Decock and Douven's basic idea is to let a conceptual space with corresponding collated Voronoi diagram define a partial model for the relevant part of the language (e.g. colour words if the space is colour space), with the collated Voronoi polygons/polyhedrons playing the part of Kamp and Partee's positive extensions, the complements of expanded collated Voronoi polygons/polyhedrons playing the part of their negative extensions, and the boundary regions associated with the prototypical regions playing the part containing the indeterminate instances. Furthermore, the elements of $\Pi(R)$ (where R is the set of prototypical regions of the space), which generate the Voronoi diagrams that together generate the collated Voronoi diagram, play the role of completions. As Decock and Douven show, and as is not difficult to see anyway, an ordinary Voronoi diagram divides the boundary region of any concept represented in the relevant space into two parts, a part containing those borderline cases that are grouped with the clear cases, and a part containing the remaining borderline cases. This kind of division of the boundary region is clearly in the spirit of Kamp and Partee, considering that Voronoi diagrams are functions of similarity rankings.

It is relatively straightforward to flesh out in this framework the idea of measuring degree of membership in terms of a measure on the set of completions. The intuitive idea behind the measure that Decock

 © Douven and Decock 2017

and Douven define is most easily grasped by considering a space with prototypical regions, all of which consist of only finitely many points. For this kind of space, the degree to which an object falls under a concept represented in the space simply equals the ratio between the number of the n-tuples that generate Voronoi diagrams in which the object is grouped with the clear cases and the total number of n-tuples. For instance, given a space S with set R of prototypical regions, each of which consists of only finitely many points, let $r_C \in R$ be the region of prototypes of concept C. Then, using the notation introduced in §2.2, the degree $M_C(x)$ to which x is a member of C is given by this function:

$$M_C(x) := \frac{\left|\left\{\vec{p} \mid x \in v(p) \wedge p \in r_C\right\}\right|}{|\Pi(R)|}.$$

Spaces whose prototypical regions contain only finitely many points may not be very realistic. But Decock and Douven show that the above simple measure generalizes swiftly to conceptual spaces some or all of whose prototypical regions consist of infinitely many points.[17]

The generalization exploits the fact that completions, now conceived of as n-tuples generating Voronoi diagrams, are themselves representable as points in a space. More exactly, given an m-dimensional conceptual space with n prototypical regions, each completion can be represented as a point in an $(m \times n)$-dimensional space. Then the degree to which an individual falls under a concept is defined to equal the volume of the set of points representing completions that group the individual with the clear instances of the concept relative to the volume of the total $(m \times n)$-dimensional space.

More formally, let S be an m-dimensional space with ordered set $R = \langle r_1, \ldots, r_n \rangle$ of prototypical regions. As mentioned, we can think of each n-tuple of prototypical points that contains exactly one point from each prototypical region — that is, of each $\vec{p} = \langle p_1, \ldots, p_n \rangle \in \Pi(R)$ — as a completion of the partial model determined by the collated Voronoi diagram on S. Because each prototypical point p_i in such an n-tuple can itself be represented as an m-tuple $\langle x_1^i, \ldots, x_m^i \rangle$ of spatial coordinates, each completion can be represented as an $(m \times n)$-tuple $\langle x_1^1, \ldots, x_m^1; \ldots; x_1^n, \ldots, x_m^n \rangle$ of real numbers. Now,

[17] It is assumed that each of these regions is *connected* in the sense that there is a continuous path (not necessarily a straight line) from any point in the region to any other point in the region, where every point lying on this path also lies in the region. The argument for this assumption is made in Douven et al. (2013).

 © Douven and Decock 2017

the crucial part is the definition of a measure over a given set of completions in terms of the volume occupied by the related set of coordinates in the space $\mathbb{R}^{m\times n}$. To this end, Decock and Douven introduce an indicator function $\mathcal{I}_T$ for each subset T of this space, the Lebesgue measure of which is defined as

$$\mu(T) := \int \mathcal{I}_T\left(\langle x_1^1, \ldots, x_m^1; \ldots; x_1^n, \ldots, x_m^n\rangle\right)\, dx_1^1 \ldots dx_m^1 \ldots dx_1^n \ldots dx_m^n$$

with $\mathbb{R}^{m\times n}$ being the domain of integration.[18] In particular, the measure of the entire set of completions is determined through the indicator function $\mathcal{I}_{\Pi(R)}$:

$$\begin{aligned}\mu(\Pi(R)) &= \int \mathcal{I}_{\Pi(R)}\left(\langle x_1^1, \ldots, x_m^1; \ldots; x_1^n, \ldots x_m^n\rangle\right)\, dx_1^1 \ldots dx_m^1 \ldots dx_1^n \ldots dx_m^n \\ &= \int_{V_{r_1}} \cdots \int_{V_{r_n}} dx_1^1 \ldots dx_m^1 \ldots dx_1^n \ldots dx_m^n \\ &= \prod_{i=1}^{n} V_{r_i}\end{aligned}$$

with V_{r_i} being the volume of region r_i. Given the plausible assumption that prototypical regions are always bounded, the measure $\mu(\Pi(R))$ will assume only finite values.

From here, it is but a small step to define the desired graded membership function. Observe that, for any point a and any concept C_i with prototypical region r_i, where δ is the distance function defined on the relevant space, the set

$$\begin{aligned}S_{a,i} := \{\langle x_1^1, \ldots, x_m^1; \ldots; x_1^n, \ldots, x_m^n\rangle \mid &\text{for all } j \neq i \text{ and all } \langle x_1^j, \ldots, x_m^j\rangle, \\ &\delta\left(a, \langle x_1^i, \ldots, x_m^i\rangle\right) < \delta\left(a, \langle x_1^j, \ldots, x_m^j\rangle\right)\}\end{aligned}$$

is the set of elements of $\Pi(R)$ that determine Voronoi diagrams that locate a in the Voronoi polygon/polyhedron associated with $\langle x_1^i, \ldots, x_m^i\rangle \in r_i$. The measure of this set is given by $\mu(S_{a,i})$. Now all

[18] One can think of the Lebesgue measure as generalizing to n-dimensional spaces the customary ways of measuring the subsets of one-, two- and three-dimensional spaces; that is to say, the Lebesgue measure generalizes the notions of length, area and volume. For mathematical details, see Temple (1971).

 © Douven and Decock 2017

one has to do to obtain a measure of the relative size of the set of completions in $S_{a,i}$ (relative to the set $\Pi(R)$, that is) is simply to define

$$\mu^*(S_{a,i}) := \frac{\mu(S_{a,i})}{\mu(\Pi(R))}$$

which is a normalized measure. The desired graded membership function is then obtained by setting $M_{C_i}(a) := \mu^*(S_{a,i})$ for all a and i.

The resulting function not only satisfies all of Kamp and Partee's adequacy requirements for graded membership functions (various formal requirements and the requirement to respect typicality rankings), but is also 'S-shaped', in the sense that as we move away from the clear cases of a concept, at some point the degree of membership starts to slowly decrease, then gradually comes to decrease at a more rapid pace, and finally, when we get close to the clear non-cases, decreases more slowly again until it reaches 0.[19] This is important in view of the fact that all extant data on people's attributions of degrees of membership seem to indicate that membership functions are S-shaped in the given sense (Hampton 2007; Douven 2016b; Douven *et al.* 2017). Moreover, Decock, Dietz and Douven (2013) show that, thus defined, the membership function can also supply an adequate semantics for comparative concepts.[20]

3 From graded membership to verities

The above account of graded membership allows us to say that an ORANGE/RED borderline case being neither quite orange nor quite red amounts formally to its being a member to some non-extreme degree (or degrees) of both the class of red things and the class of orange things. That an ORANGE/RED borderline case is neither quite orange nor quite red can equally be stated in terms of truth: it is neither quite true that the object is orange nor quite true that it is red. Combining these two facts makes for a straightforward definition of a notion of graded truth (or verity) for atomic sentences: the sentence $\ulcorner \mathrm{F}(a) \urcorner$ is true to degree x iff a belongs to the class of Fs to degree x, or formally, $\mathrm{Ver}(\mathrm{F}(a)) = M_{\mathrm{F}}(a) = x$.[21] But how can this definition be extended

[19] See Decock and Douven (2014) for various illustrations of the 'S-shapedness' of membership functions.

[20] For an alternative model of comparative concepts that also builds on Douven et al.'s (2013) proposal, see Dietz (2013).

[21] In '$M_{\mathrm{F}}(a)$', the subscript stands for the concept that is the referent of the predicate F.

 © Douven and Decock 2017

to the rest of the language? Specifically, how can it be extended such that the Ver operator is governed by the probability axioms?

The intuitive (but rough) idea underlying the account of graded membership described in §2.3 is that something falls under a given concept to a degree that corresponds to the 'proportion' of simple Voronoi diagrams that jointly constitute the collated Voronoi diagram on the space in which the concept is represented and that locate the thing in the Voronoi polygon/polyhedron associated with the concept's prototypes. This was seen to be a geometrization of part of Kamp and Partee's supervaluationist approach to vagueness, letting simple Voronoi diagrams play the role of completions.

To complete Kamp and Partee's programme in the same geometrical setting, and to do this in their spirit, one would like — at a minimum, and to a first approximation — $\ulcorner \overline{\mathrm{F}(a)} \urcorner$ to be true to a degree corresponding to the 'proportion' of Voronoi diagrams that do *not* locate a in the Voronoi polygon/polyhedron associated with the F-prototypes; $\ulcorner \mathrm{F}_1(a) \,\&\, \mathrm{F}_2(b) \urcorner$ to be true to a degree corresponding to the 'proportion' of Voronoi diagrams that locate a in the Voronoi polygon/polyhedron associated with the F_1 prototypes *and* locate b in the Voronoi polygon/polyhedron associated with the F_2 prototypes; and $\ulcorner \mathrm{F}_1(a) \vee \mathrm{F}_2(b) \urcorner$ to be true to a degree corresponding to the 'proportion' of Voronoi diagrams that locate a in the Voronoi polygon/polyhedron associated with the F_1 prototypes *or* locate b in the Voronoi polygon/polyhedron associated with the F_2 prototypes. (Following Edgington 1997a, p. 310, we assume that there are only finitely many objects in the universe of discourse, all of which are named, so that existential quantification can be defined in terms of disjunction, and universal quantification in terms of conjunction. The indicative conditional operator requires special attention; see below.)

One reason this can only be a first approximation is that the concepts to which F_1 and F_2 refer need not be representable in the same space, and collated Voronoi diagrams are defined only for single spaces. Another is that we cannot literally speak of proportions, given that collated Voronoi diagrams are typically composed of infinitely many (simple) Voronoi diagrams. But for purposes of illustration, we ignore these complications for the moment, and consider the kind of case in which the relevant concepts are representable in the same space and the collated Voronoi diagram defined on that space is composed of finitely many simple Voronoi diagrams.

Consider, for example, the space-cum-collated Voronoi diagram depicted in Fig. 3. The sixteen simple Voronoi diagrams that make

 © Douven and Decock 2017

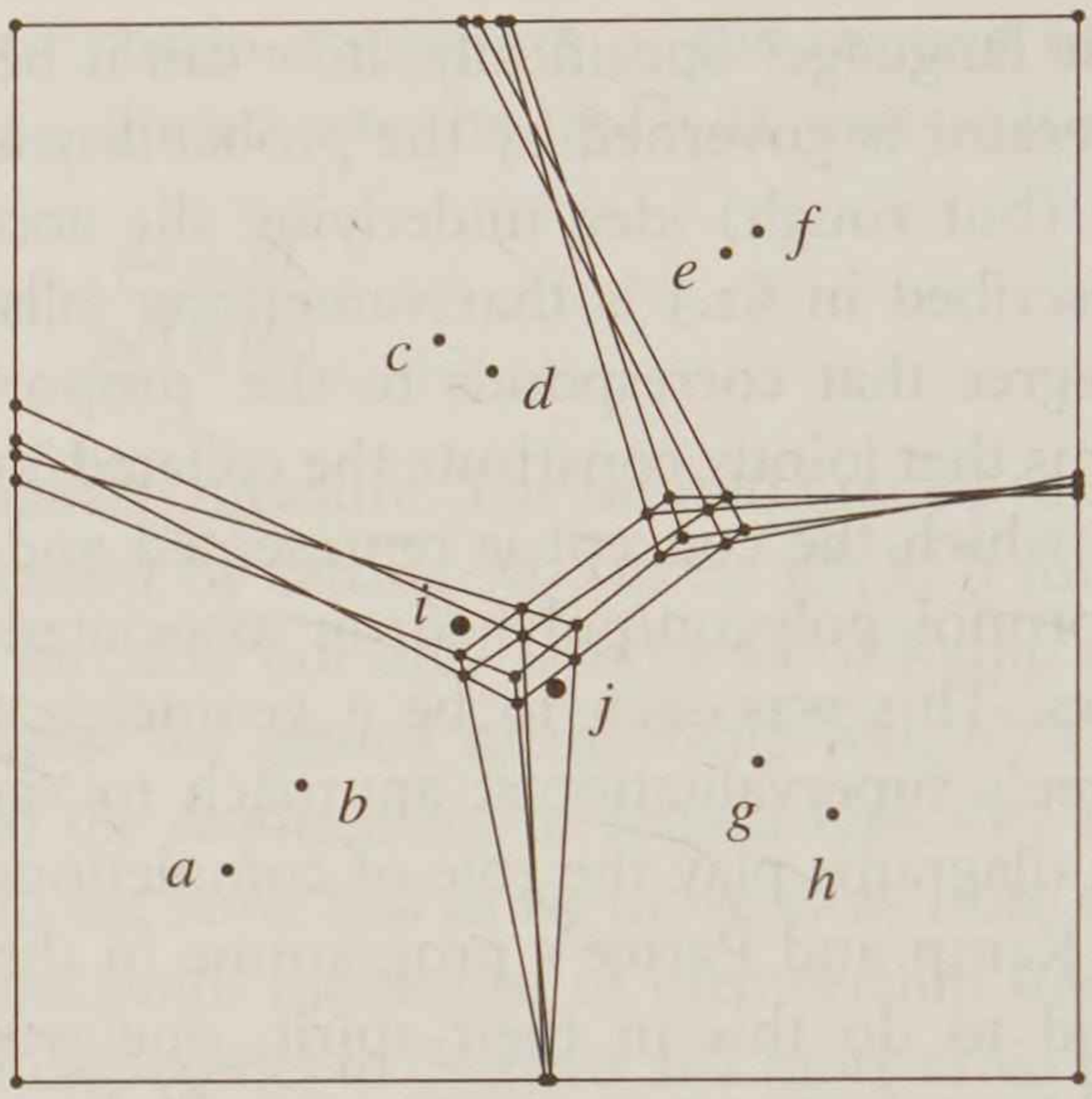

Figure 3: A collated Voronoi diagram of a space with prototypical areas each consisting of only two points and with designated points i and j

up this collated diagram are presented individually in Fig. 4. Both figures are taken from Decock and Douven (2014), where they are used to introduce the idea behind the measure of graded membership advocated in that paper. As noted there, and as one can readily verify, $M_{\{a,b\}}(i) = M_{\{c,d\}}(i) = 1/2$; $M_{\{a,b\}}(j) = 1/4$; $M_{\{g,h\}}(j) = 3/4$; and $M_{\{g,h\}}(i) = M_{\{e,f\}}(i) = M_{\{c,d\}}(j) = M_{\{e,f\}}(j) = 0$ (where $M_{\{a,b\}}$ denotes the graded membership function for the concept with prototypical region $\{a, b\}$, etc.).

Now let F_1 refer to the concept with prototypical region $\{a, b\}$ and F_2 refer to the concept with prototypical region $\{g, h\}$. According to the above (rough) proposal, we have, for instance, $\mathrm{Ver}(F_1(i)) = M_{F_1}(i) = 1/2$ and $\mathrm{Ver}(F_2(j)) = M_{F_2}(j) = 3/4$. According to the same proposal, $\mathrm{Ver}(F_1(i) \,\&\, F_2(j))$ should equal the proportion of simple diagrams that locate i with $\{a, b\}$ and j with $\{g, h\}$. Inspection of Fig. 4 shows that this is four out of sixteen, so $\mathrm{Ver}(F_1(i) \,\&\, F_2(j)) = 1/4$. What is $\mathrm{Ver}(F_1(i) \vee F_2(j))$? Again, the idea is that this should equal the proportion of simple diagrams that locate i with $\{a, b\}$ or j with $\{g, h\}$ (or both). We see that this is sixteen out of sixteen, so $\mathrm{Ver}(F_1(i) \vee F_2(j)) = 1$. According to the second verity axiom in §1 above, this should equal $\mathrm{Ver}(F_1(i)) + \mathrm{Ver}(F_2(j)) - \mathrm{Ver}(F_1(i) \,\&\, F_2(j))$, which it does, since $1/2 + 3/4 - 1/4 = 1$.

 © Douven and Decock 2017

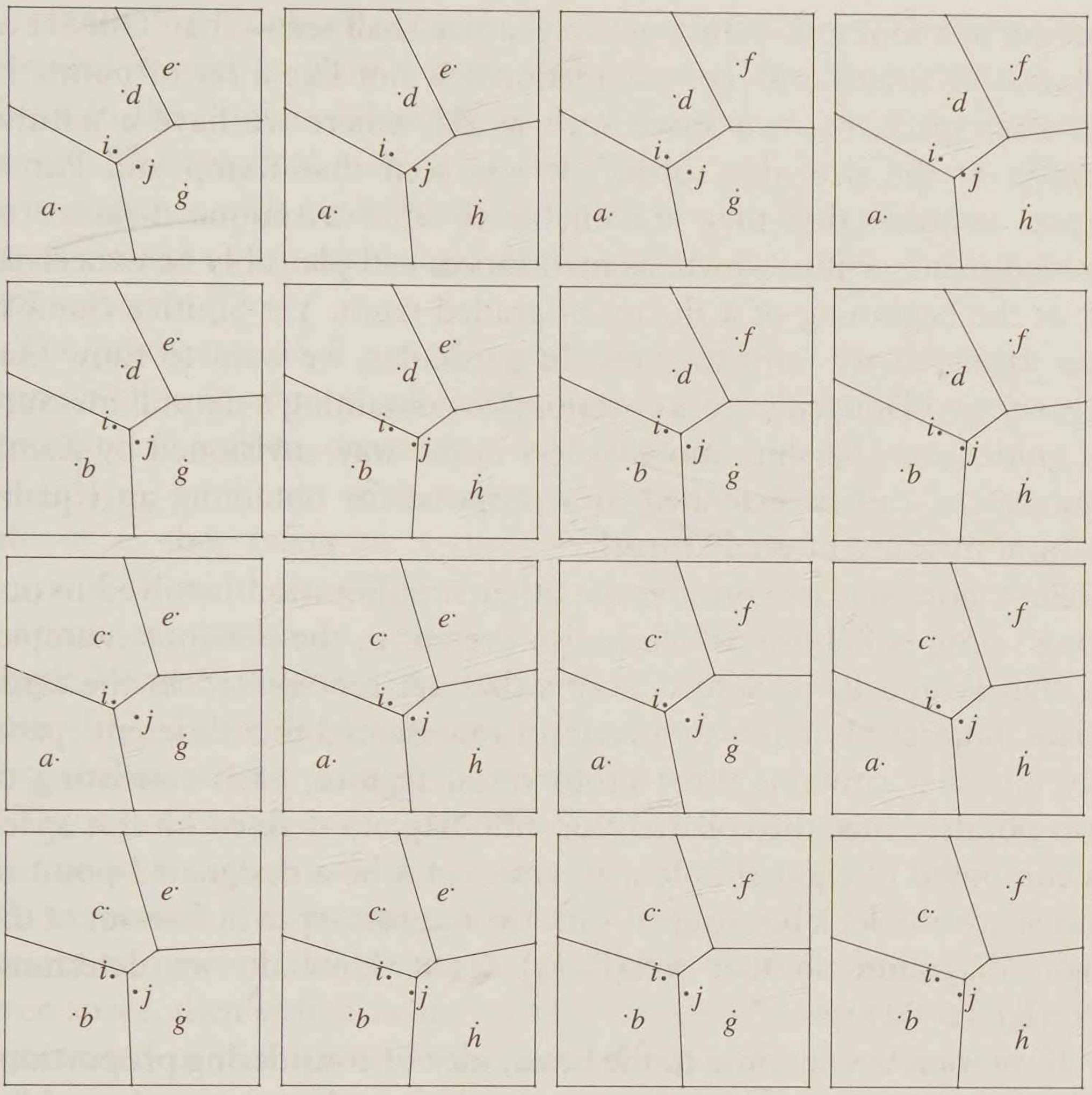

Figure 4: The sixteen Voronoi diagrams that make up the collated Voronoi diagram shown in Fig. 3

This is entirely unsurprising. After all, we are considering proportions, and proportions are nothing but relative frequencies, which are known to obey the probability axioms. The question is whether there is a natural generalization to more realistic spaces — notably, spaces whose prototypical regions consist of infinitely many points — such that the probability axioms are satisfied.

The step from finite to infinite is far from trivial, although — excepting Kamp and Partee (1995) — previous authors who aimed to develop a degree-theoretic account of vagueness in a supervaluationist framework seem not to have realized this. Indeed, Lewis (1970), Kamp (1975) and Akiba (2004) do not even consider this step. Smith (2008, p. 89) is right in regarding this as an important lacuna. In his discussion of these earlier proposals, he even goes so far as to claim that 'in the infinite case, there is no non-arbitrary, natural choice of measure',

 © Douven and Decock 2017

adding in a footnote — instructively, as we shall see — that 'Our set of admissible extensions [i.e. completions] is not like a set of points in physical space, or in a space such as $\mathbb{R}^n$, where we have a natural notion of the size of a subset'. It was seen that Kamp and Partee (1995) concede that they are unable to offer a unique measure of graded membership, which, as mentioned, can plausibly be conceived of as the beginning of a theory of graded truth. Yet Smith's claim is too strong, as we want to show. In particular, we want to show that Decock and Douven's (2014) proposal for obtaining a natural measure of graded membership, more or less in the way envisioned by Kamp and Partee, can be extended to a proposal for obtaining an equally natural measure of graded truth.

First, however, we consider the other simplification involved in our rough proposal above, which is also present in the previous example, in that F_1 and F_2 refer to concepts that are represented in the same space. Suppose F_3 refers to a concept represented in a different space, say one that contains three prototypical regions, each consisting of two points. Thus, the collated Voronoi diagram defined on this space is composed of eight simple diagrams. Let k be a designated point in this space and let k be grouped with the F_3 prototypes in five out of the eight diagrams, so that $\mathrm{Ver}(F_3(k)) = 5/8$. How do we determine $\mathrm{Ver}(F_1(i) \,\&\, F_3(k))$?

If we want to stay close to the broad idea of considering proportions of simple diagrams, the obvious answer is to look at the *product* of the set of simple diagrams defined on the space in which the concept that F_1 refers to is represented and the set of simple diagrams defined on the space in which the concept that F_3 refers to is represented, and then determine how many of the $16 \times 8 = 128$ pairs of diagrams in the resulting set are such that the first element of the pair locates i with the F_1 prototypes and the second element locates k with the F_3 prototypes. One can easily check that this holds for 40 pairs. As a result, $\mathrm{Ver}(F_1(i) \,\&\, F_3(k)) = 40/128 = 5/16$. One can just as easily check that $\mathrm{Ver}(F_1(i) \,\&\, F_3(k)) = 104/128$, which equals $\mathrm{Ver}(F_1(i)) + \mathrm{Ver}(F_3(k)) - \mathrm{Ver}(F_1(i) \,\&\, F_3(k)) = 1/2 + 5/8 - 5/16$. Again, this is not surprising, given that we are still considering proportions of elements of a finite set.

To further generalize and render precise these preliminary remarks, we define a verity function for a language $\mathcal{L}$ whose logical vocabulary consists of the negation operator, the conjunction operator, and the disjunction operator, and whose non-logical vocabulary consists of finitely many monadic predicates and finitely many individual

 © Douven and Decock 2017

constants. We think of $\mathcal{L}$ as an *interpreted* language, and assume its intended model to have a finite domain that is fully named (every object in the domain is named by an individual constant in the language). The practical motivation for this assumption is that it allows us to dispense with quantification, which would otherwise raise sticky issues related to countable additivity that are orthogonal to our present concerns.[22]

The predicates of $\mathcal{L}$ are all assumed to refer to concepts that are representable in conceptual spaces, where these spaces may now have prototypical regions containing infinitely many prototypical points. Let $\{S^k\}_{k\leq w}$ be the set of conceptual spaces such that there is a predicate of $\mathcal{L}$ that refers to a concept representable in S^k, for some $k \leq w \in \mathbb{N}$; so, these are all the spaces we need for representing the concepts expressible in $\mathcal{L}$. As previously defined, $\Pi(R^k)$ is the set of simple Voronoi diagrams generated by the prototypical regions in S^k. Thus, $\prod_{k=1}^{w} \Pi(R^k)$ is the product of all the sets of simple Voronoi diagrams associated with the elements of $\{S^k\}_{k\leq w}$.

The set of sentences of $\mathcal{L}$ is defined in the usual way, except that we assume each sentence to be semantically well-formed, in the sense of this expression that is now common in linguistics.[23] In terms of the present framework, this means that if an object is not representable in a given space, then sentences stating that the object falls under a concept represented in that space, or sentences that contain a subformula that makes such a statement, do not belong to the language. For example, a melody is not measurable on any of the axes of colour space and so is not representable in that space. Accordingly, sentences attributing colour properties to melodies (in a literal sense) are excluded from $\mathcal{L}$.

Recall Smith's remark, made as a comment on earlier degree-theoretic supervaluationist proposals, that the key problem with finding a unique measure of graded truth as a function of sets of completions is that completions do not form a set of points in a space like $\mathbb{R}^n$. The same problem prevented Kamp and Partee from arriving at a unique graded membership function. What helped Decock and Douven overcome this problem was the fact that if completions can be interpreted geometrically, they can be thought of as sets of coordinates, and coordinates do form a space for which there exists a standard measure (the Lebesgue measure). In basically the same way, one obtains a measure of graded truth. The only difference is that now we must

[22] See, for example, Earman (1992, pp. 36–7, 60–1).

[23] See, for instance, Lyons (1995, ch. 5).

 © Douven and Decock 2017

consider not separate completions — simple Voronoi diagrams — but elements of the product set $\prod_{k=1}^{w} \Pi(R^k)$, which are sequences of completions. These sequences can again be represented in a space of the right dimensionality, and the measure of graded truth is then given by the Lebesgue measure on that space.

To make this formal, we first define a mapping $[\![\cdot]\!]$ of sentences of $\mathcal{L}$ onto subsets of $\prod_{k=1}^{w} \Pi(R^k)$. Let F_i^o refer to the concept that is represented by the collated Voronoi polygon or polyhedron associated with r_i^o, the ith prototypical region in S^o. Then for all predicates F_i^o and all individual constants a such that $\ulcorner \mathrm{F}_i^o(a) \urcorner \in \mathcal{L}$,

$$[\![\mathrm{F}_i^o(a)]\!] := \{\vec{x} \in \prod_{k=1}^{w} \Pi(R^k) \mid \text{for all } j \neq i \text{ and all } \langle x_1^{j,o}, \ldots, x_{m^o}^{j,o} \rangle \in r_j^o,$$

$$\delta\left(a, \langle x_1^{i,o}, \ldots, x_{m^o}^{i,o} \rangle\right) < \delta\left(a, \langle x_1^{j,o}, \ldots, x_{m^o}^{j,o} \rangle\right)\}$$

Furthermore, for all $\phi, \psi \in \mathcal{L}$,

$$\begin{aligned} [\![\phi \wedge \psi]\!] &:= [\![\phi]\!] \cap [\![\psi]\!]; \\ [\![\phi \vee \psi]\!] &:= [\![\phi]\!] \cup [\![\psi]\!]; \\ [\![\bar{\phi}]\!] &:= \left(\prod_{k=1}^{w} \Pi(R^k)\right) \backslash [\![\phi]\!] \end{aligned}$$

To explain the atomic clause, note that the elements of $\prod_{k=1}^{w} \Pi(R^k)$ are all of the form $\langle \langle x_1^{1,1}, \ldots, x_{m^1}^{1,1}, \ldots, x_1^{n^1,1}, \ldots, x_{m^1}^{n^1,1} \rangle, \ldots, \langle x_1^{1,w}, \ldots, x_{m^w}^{1,w}, \ldots, x_1^{n^w,w}, \ldots, x_{m^w}^{n^w,w} \rangle \rangle$, with m^o the number of dimensions of space S^o, n^o the number of prototypical regions in S^o, and $x_j^{i,o}$ the jth coordinate of a selected point in r_i^o. That is to say, each element is a sequence of sequences, $s_1, \ldots, s_w$, where s_o is a string of $m^o \times n^o$ real numbers, the first m^o of which are the coordinates of a point from prototypical region r_1^o, the second m^o of which are the coordinates of a point from r_2^o, and so on. Together, these coordinates generate a simple Voronoi diagram on S^o. Then to determine whether an element $\vec{x} \in \prod_{k=1}^{w} \Pi(R^k)$ is contained in the set $[\![F_i^o(a)]\!]$, we only have to look at the oth sequence in $\vec{x}$ and see whether the point a is closer to the prototypical point in r_i^o whose coordinates are given by the oth element of $\vec{x}$ than to the prototypical points whose coordinates are given by the other elements of $\vec{x}$.

It may help to note that, intuitively, one can think of the elements of $[\![\phi]\!]$ as those elements of $\prod_{k=1}^{w} \Pi(R^k)$ that 'precisify' the predicates

 © Douven and Decock 2017

of $\mathcal{L}$ in such a way as to make ϕ come out clearly true. It is also to be noted, however, that the definition does not employ any semantic notions, and so does not require that an account of truth or clear truth (let alone graded truth) already be in place.

To come to our definition of verity, the final observation to be made is that we can uniquely identify each element $\vec{x} \in \prod_{k=1}^{w} \Pi(R^k)$ with the sequence of numbers we obtain by deleting in $\vec{x}$ the inner angle brackets — thus turning a sequence of sequences of real numbers into a sequence of real numbers — and that the result of that operation can be represented by a point in the space $\mathbb{R}^{m^1 \times n^1 \times \cdots \times m^w \times n^w}$ (for brevity, call this space $\mathcal{S}$). Hence, each $\vec{x} \in \prod_{k=1}^{w} \Pi(R^k)$ can be identified with a point in $\mathcal{S}$, and each set of elements of $\prod_{k=1}^{w} \Pi(R^k)$ with a region in $\mathcal{S}$. Let $\mathcal{I}_T^+$ be the indicator function for subset $T \in \mathcal{S}$, and let μ^+ be the Lebesgue measure on $\mathcal{S}$. Thus, $\mu^+(T)$ is the measure of T and, in particular, $\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)$ is the measure of the entire product space, which, as one can easily verify, equals $\prod_{k=1}^{w} \prod_{j=1}^{n^k} V_{r_j^k}$. Then, for $\phi \in \mathcal{L}$, the verity of ϕ is given by

$$\mathrm{Ver}(\phi) := \frac{\mu^+([\![\phi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = \frac{\mu^+([\![\phi]\!])}{\prod_{k=1}^{w} \prod_{j=1}^{n^k} V_{r_j^k}}.$$

Given the previous definitions, it holds trivially for all atomic sentences $\ulcorner \mathrm{F}_i^k(a) \urcorner \in \mathcal{L}$ that $\mathrm{Ver}\left(\mathrm{F}_i^k(a)\right) = M_{\mathrm{F}_i^k}(a)$, where $M_{\mathrm{F}_i^k}$ is the graded membership function of the concept referred to by F_i^k. It does not follow from this that verities are probabilities. Nor does this follow simply from the fact that Ver is by construction a normalized Lebesgue measure function, which is formally a probability measure. For instance, we could have defined both $[\![\phi \wedge \psi]\!]$ and $[\![\phi \vee \psi]\!]$ as $[\![\phi]\!] \cup [\![\psi]\!]$, in which case we could hardly have expected verities to have probabilistic structure (we do not have in general that $\Pr(\phi \wedge \psi) = \Pr(\phi \vee \psi)$). Nevertheless, it is rather straightforward to prove

Theorem 3.1. Verities are formally probabilities.

(See the Appendix for a proof.)

We have a number of comments on the foregoing. The first concerns the fact that the definition of verity is limited to predicates that refer to concepts that can be identified with a region in a conceptual

 © Douven and Decock 2017

space. How much of a restriction this is depends on how general the conceptual spaces approach can be made to be, which — as mentioned in §2.1 — is largely an open question. As also mentioned there, however, the sorites paradox is typically stated in terms of a predicate referring to a perceptual concept — typical examples of the paradox exploit the fact that perceptually indiscriminable differences can add up to a perceptually discriminable difference — and the conceptual spaces approach is generally regarded as being very successful in handling such concepts. So, at a minimum, we are considering an important fragment of the language, possibly even precisely the fragment for which the notion of verity is relevant.

Second, the definition is restricted to monadic predicates. Gärdenfors (2014, chs. 2, 11) mentions that relational concepts can also be represented in the conceptual spaces framework, that is — he suggests — in product spaces. The evidence supporting this suggestion may not be quite as strong as the evidence supporting the use of conceptual spaces to represent non-relational concepts. Nevertheless, if the suggestion is correct, the definition of the verity operator can easily be generalized to polyadic predicates.[24]

Third, we saw that Edgington left open the question of the verities of indicative conditionals, although she deemed it tempting to equate

[24] It would be a mistake to think that, in the context of a discussion of the sorites paradox, polyadic predicates are not really important. For although the sorites paradox is usually stated in terms of a monadic predicate like 'being bald', 'being rich', 'being red' or 'being a heap', this is not an essential feature of the paradox. Keefe (2000, p. 14) mentions the predicate 'is a friend of' as one permitting of borderline cases. It may even be easier to give a version of the sorites paradox for 'is badly out of tune with'. For instance, consider a series of 50 violins, numbered 1 through 50, such that for each violin, it holds that its strings are well-tuned relative to one another. However, whereas the A string of violin number 1 is 440 Hz, the A string of violin number 2 is 489 Hz, that of violin number 3 441 Hz, that of violin number 4 488 Hz, and so on; so between the A strings of the first and second violins there is a difference of 49 Hz, between the A strings of the second and third violins there is a difference of 48 Hz, and so on, till there is a difference of only 1 Hz between the A strings of violins number 49 and 50. A difference of 1 Hz is in this range exceedingly difficult to hear, even for the trained ear. On the other hand, the difference of 49 Hz makes violins number 1 and 2 badly out of tune with each other. Now consider the following argument:

P_0 Violins numbers 1 and 2 are badly out of tune.
P_1 If violins numbers 1 and 2 are badly out of tune, so are violins numbers 2 and 3.
P_2 If violins numbers 2 and 3 are badly out of tune, so are violins numbers 3 and 4.
⋮ ⋮
P_{48} If violins numbers 48 and 49 are badly out of tune, so are violins numbers 49 and 50.
??∴ Violins numbers 49 and 50 are badly out of tune.

In view of the stated fact that a difference of 1 Hz (in the given range) is very hard to detect, each of $P_1, \ldots P_{48}$ seems true. In view of the same fact, however, the conclusion seems false.

 © Douven and Decock 2017

them with the corresponding conditional verities. She did not commit to the equation, we hypothesized, in view of the triviality arguments that have been levelled against Stalnaker's Hypothesis. As we also noted, however, these arguments all make assumptions beyond Stalnaker's Hypothesis, and it is easy to see that on our proposal of what verities are, these additional assumptions do not make sense.

We mentioned that many triviality arguments assume that there are at least two different probability functions. Obviously there is but one verity function, determined by the structure of conceptual reality. For similar reasons, the assumptions underlying most of the other triviality arguments have no natural interpretation in terms of verities. While we understand what is meant when we are asked to assume that probability is transferred from one proposition to another — we might have believed one proposition more strongly and another less strongly than we actually do — verities are ultimately fixed by the structure of conceptual reality, which, as explained in §2.1, is grounded in people's similarity judgements. Colour space is what it is; we can squeeze it a bit here, and stretch it a bit there, but then it is no longer colour space, no longer a representation of the actual data concerning judged similarities. That is enough to render, for instance, Hájek's 1994 argument inapplicable. Similarly, it is not possible to restrict the range of the verity operator to some finite subset of [0, 1], as would be a minimal requirement for Hájek's 1989 argument to apply. To appreciate this, it suffices to observe that, for instance, the graded membership function of any colour concept already takes all values in the [0, 1] interval. In sum, for all anyone has shown, we can adopt (EQ) without risking triviality.[25]

[25] Admittedly, there is a triviality argument due to Stalnaker (1976) that our interpretation of verities does not help to block. However, this argument is also rather weak and, at least in the context of a discussion of verities, not really worrying. The argument crucially relies on the assumption that indicative conditionals are governed by Stalnaker's logic C2, which one may find objectionable (see, for example, van Fraassen 1976). More importantly, it follows from results to be found in van Fraassen (1976) that even if we do assume that C2 is the logic of indicative conditionals, (EQ) still does not engender triviality as long as one is willing to limit the scope of the principle somewhat: specifically, to restrict it to simple conditionals (conditionals whose antecedent and consequent are not themselves conditional in form), simple right-nested conditionals (conditionals whose antecedent is not conditional in form and whose consequent is a simple conditional), and simple left-nested conditionals (conditionals whose antecedent is a simple conditional and whose consequent is not conditional in form). This is very mild restriction if, as some hold, we have difficulty making sense of iterated conditionals, or at least of iterated conditionals that cannot be reduced to non-iterated conditionals (see Jackson 1987, pp. 127–37; Edgington 1995, pp. 382 ff.). Even if these theorists are wrong (see Douven 2016a, ch. 2), restricting as indicated the scope of (EQ) — which concerns verities, and not probabilities — seems harmless in so far as our interest in the principle arises from the fact that we seem free to interpret the conditionals in standard sorites arguments as

 © Douven and Decock 2017

Fourth, Smith (2008) complains that degree supervaluationism must lack semantic uniformity, in that atomic sentences and compound sentences receive their degrees of truth 'in fundamentally different ways':

> [S]aying that 'Bob is bald' has a degree of truth of 0.3, or that 'John is tall' has a degree of truth of 0.5, really means something quite different from saying that 'Bob is bald and John is tall' has a degree of truth of 0.25. The first sentence has an intermediate degree of truth because as a matter of fact, out there in the world, Bob neither completely possesses the property of baldness, nor completely fails to possess it—he possesses it to degree 0.3. Similarly for the second sentence. But the third sentence has an intermediate degree of truth for a completely different reason: because one quarter of the [completions] assign it the value 1. (Smith 2008, pp. 90–1)

In Smith's view, this difference is 'arbitrary, unmotivated, and downright odd' (2008, p. 91). He considers the possibility of fixing the problem by letting the degrees of truth of atomic sentences be given in the exact same way as the degrees of truth of compound sentences are given, that is, by letting the degree of truth of an atomic sentence equal the measure of the completions on which it comes out true. But this gets things backwards, he thinks, for the degree of truth of 'Bob is bald' should really depend on the degree to which Bob possesses the property of baldness, and should have nothing to do with which completions do or do not make the sentence true (2008, p. 92). Also, it might lead to bizarre conflicts. What if Bob possesses the property of being bald to a degree of 0.3 whereas 'Bob is bald' comes out true on a set of completions of measure 0.4? According to Smith, it should be clear that in this case 'Bob is bald' is true to a degree of 0.3, not 0.4.

We share this judgement, and also believe that the general point is spot on as a critique of earlier degree-theoretic forms of supervaluationism. However, that any such form *must* fail to be semantically uniform or else assign degrees of truth to atomic sentences in ways that are intuitively wrong, as Smith suggests, is incorrect. Our own proposal is a case in point: atomic and compound sentences receive their degrees of truth in exactly the same way, by measuring the size of the set of completions that make them come out true. In view of Decock and Douven's (2014) proposal of measuring graded membership, this ensures that, for any object *a* in our domain and any property F referred to by a predicate

---

indicatives. The restricted principle then still answers our question of what the verities of those conditionals are, given that, to the best of our knowledge, no such argument involves any iterated conditionals (let alone conditionals of greater complexity than those admitted by the restricted principle).

© Douven and Decock 2017

in the vocabulary of $\mathcal{L}$, the degree to which *a* possesses the property of F-ness equals the degree of truth of $\ulcorner \mathrm{F}(a) \urcorner$. This is ensured not simply by fiat, but because the completions that determine the degree to which $\ulcorner \mathrm{F}(a) \urcorner$ is true are at the same time the completions that determine the degree to which *a* possesses the property of F-ness.

Finally, there is the issue of higher-order vagueness. As we saw, in light of this issue Edgington tried to downplay the precision of verities as being artefactual. As we also saw, however, Keefe questioned the legitimacy of this move, absent a principled basis for determining which elements of a theory are artefactual and which represent features of reality. Independently of Keefe's point, it is not open for us to dismiss even the slightest difference in verity as being artefactual.[26] In our account, slight differences in verity ultimately derive from slight differences in location in conceptual space. But then we do face the challenge of explaining how vagueness—apparent *lack* in precision—is compatible with the high degree of precision that comes with our account. In particular, in our account the clear instances of, say, 'red' are sharply delineated from the borderline cases of that predicate, which in turn are sharply delineated from its clear non-instances: the clear instances are those objects to which the predicate applies with verity 1; the clear non-instances are those objects to which the predicate applies with verity 0; and the borderline cases are those objects to which the predicate applies with some intermediate verity. Thus, no sense can be made of an object being *borderline* borderline red, let alone borderline borderline borderline red, and so on.

However, the point to be stressed is that there really is no incompatibility between the precision that our account offers and the phenomena of higher-order vagueness. In fact, we believe that Smith goes a long way toward showing this when he writes, 'The demand for higher and higher orders of borderline cases is just the demand for a gradual transition from the cases where the predicate clearly applies to the cases where it clearly does not apply—rather than a transition

[26] Note that we are not committed to conceiving of these differences as being *absolute*. There is evidence suggesting that the selection of a conceptual space may be relative to an observer as well as to a point in time, for instance, that the exact geometrical features of colour space may depend on the observer and the viewing conditions the observer is in (see Fairchild 1998, pp. 168–9; see Nosofsky 1986 for the more general point).

 © Douven and Decock 2017

marked by a series of jolts, whether one big jolt or two smaller jolts or four even smaller jolts, and so on' (2008, p. 173).[27] For gradual transitions are precisely what we get from our account. If we traverse colour space in a straight line from the clear cases of RED to the clear cases of ORANGE, then the first borderline cases we come across are ones to which the predicate 'red' applies with a verity *exceedingly* close to 1. As we move on, the verity with which the predicate applies to the cases we encounter will taper off — gradually. Surely that explains why there is nothing in the way we experience the transition from red to orange that signals the presence of categorical shifts.

Yet this does not seem to fully explain higher-order vagueness. For it remains the case that, on our account, there are sharp boundaries between the clear cases of a concept, the clear non-cases, and the borderline cases. But then why can we not, even if only with effort, pinpoint the place in a soritical series where the relevant predicate ceases to apply with perfect verity? That the transition is gradual may explain why we do not experience any abruptness in going through the series, but not why — as it appears — we cannot locate the line separating perfect and non-perfect applicability of the predicate, *try as we may*, using *any* method we like. An equally important question that is not answered by Smith's observation — correct and important by itself — is the question of where the boundaries come from. What in the world determines that the boundary line lies between, say, patch number 6 and patch number 7, and not between any other pair of adjacent patches?[28]

The first question points to the fact that an additional factor — next to the smoothness of the transition from clear cases to borderline cases and from these to clear non-cases — *is* involved in higher-order vagueness, to wit, the imprecision of psychological measures, such as the metrics defined on conceptual spaces, as emphasized by Gärdenfors (2000, p. 89; 2014, ch. 2). While conceptual reality determines precise boundaries between the different types of cases, this reality is not surveyable with infinite precision. The mind's eye has limited discriminatory capacities. In particular, it is not discriminating enough to be able to sharply locate the boundary between the restricted collated

[27] Also in this vein, see Hampton (2007).

[28] In Smith's own account, this is a primitive fact, reflecting different degrees of property instantiation in the world. For a detailed comparison of Smith's 'realist' degree-theoretic account of vagueness, which locates vagueness in the world, with the present account, in which vagueness primarily arises from features of our representational system, see Decock and Douven (2017).

 © Douven and Decock 2017

Voronoi polygon/polyhedron associated with a prototypical region and the boundary region associated with the same prototypical region.

The second question is generally regarded as posing a major challenge to epistemicist accounts of vagueness, which assume every predicate to have sharp application conditions and attribute vagueness to our irresolvable ignorance of the boundary between correct and incorrect applications of the predicate (see Sorensen 1988 and Williamson 1994). The advocates of these accounts have been virtually unable to explain to anyone's satisfaction the putative existence of these boundaries. Most critics agree with Sainsbury (1990, p. 255) that 'no sharp cut-off of the shadow of vagueness is marked in our linguistic practice, so to attribute it to the predicate is to misdescribe it'.

On our account, the second question has an obvious answer. The sharp cut-offs associated with a predicate are not determined by linguistic practice. That much we grant Sainsbury. It is a mistake, however, to assume — as Sainsbury apparently does — that, if anything, then only linguistic practice could fix sharp boundaries. Indeed, our account gives the lie to that assumption by showing how instead the structure of conceptual reality gives rise to such boundaries. For instance, in colour space, the cut-off between the clear instances of 'red' and the borderline cases of that predicate is determined by the geometrical and topological structure of colour space, most notably, by the locations of the prototypical regions of RED and of the colour concepts which surround it in colour space (like ORANGE, PINK, and PURPLE). Jointly, those facts determine that the dividing line between the red patches and the borderline patches lies at one precise point in the soritical series and not at any other.[29, 30]

[29] We thus arrive at a hybrid account of vagueness that bears some resemblance to MacFarlane's (2010) fuzzy epistemicism, which combines a fuzzy logic framework for first-order vagueness with an epistemicist account for higher-order vagueness. Hybrid accounts may raise the question of why different orders of vagueness require different explanations. In our view, there is no a priori reason to think that they do require different explanations; nor, however, is there an a priori reason to think they could not have different explanations. In a way, we followed a bottom-up approach to vagueness, which started not by asking questions about vagueness but by asking questions about conceptualization (in previous work). This approach led us to embrace a theory of concepts that, as it happens, implies a hybrid account of vagueness. That is the a posteriori reason that we can offer for explaining first- and higher-order vagueness along different lines. (Thanks to Richard Dietz here.)

[30] It will do no harm to reiterate that our background theory of conceptualization makes talk of conceptual structure entirely unmysterious: given this theory, conceptual structure is fully determined by our similarity judgements, together with our judgements of which items

 © Douven and Decock 2017

In summary, we have provided an account of verities in which these emanate from the geometry of concepts. As we have seen, this account allows us to answer questions concerning the verities of indicative conditionals that otherwise might have kept us from endorsing what appears to be the most natural proposal on this point. We have further seen that the account has no special difficulty accommodating phenomena of higher-order vagueness.

To forestall misunderstanding, we have not shown, nor did we have the intention to show, what verities *must* be if they are to play the role that Edgington wants them to play, and in particular if they are to obey the probability axioms. Rather, we have shown what verities *may* be, and have provided theoretical reasons that make it somewhat plausible, at least, to hold that we have shown what verities *are*. As a further general comment on what we hope to have accomplished, we note that nothing in the foregoing vindicates Edgington's approach to vagueness and the sorites paradox by showing this approach to be superior to rival ones, or even just superior to rival degree-theoretic approaches. Our aim was to provide foundations for an account that we find to offer a particularly attractive solution to the sorites paradox, and we have tried to do so by anchoring this account in recent scientific work on conceptualization.[31]

Appendix

This appendix presents the proof of

Theorem 3.1. Verities are formally probabilities, that is, the operator Ver obeys these axioms:

$$(1)\quad 0 = \mathrm{Ver}(\phi \wedge \bar{\phi}) \leq \mathrm{Ver}(\phi) \leq \mathrm{Ver}(\phi \vee \bar{\phi}) = 1;$$

$$(2)\quad \mathrm{Ver}(\phi \vee \psi) = \mathrm{Ver}(\phi) + \mathrm{Ver}(\psi) - \mathrm{Ver}(\phi \wedge \psi).$$

strike us as the most representative or typical of a given concept; there is nothing esoteric about either type of judgement.

[31] For helpful comments on previous versions of this paper we are indebted to Thomas Baldwin, Richard Dietz, Nick Jones, Nick Smith, Christopher von Bülow, and two anonymous referees. A version of this paper was presented at the *Probability and Vagueness* conference at the University of Tokyo. We thank Richard Dietz for the invitation, and the audience for stimulating questions and remarks.

 © Douven and Decock 2017

Proof. To show that Ver obeys Axiom 1, it suffices to observe that, by the definitions of $[\![\cdot]\!]$ and Ver, the following hold for all $\phi \in \mathcal{L}$. First,

$$\mathrm{Ver}(\phi \wedge \bar{\phi}) = \frac{\mu^+([\![\phi \wedge \bar{\phi}]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = \frac{\mu^+([\![\phi]\!] \cap [\![\bar{\phi}]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} =$$

$$\frac{\mu^+\left(\left([\![\phi]\!] \cap \prod_{k=1}^{w} \Pi(R^k)\right) \backslash [\![\phi]\!]\right)}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = \frac{\mu^+(\varnothing)}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = 0$$

Second,

$$\mathrm{Ver}(\phi \vee \bar{\phi}) = \frac{\mu^+([\![\phi \vee \bar{\phi}]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = \frac{\mu^+([\![\phi]\!] \cup [\![\bar{\phi}]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} =$$

$$\frac{\mu^+\left(\left([\![\phi]\!] \cup \prod_{k=1}^{w} \Pi(R^k)\right) \backslash [\![\phi]\!]\right)}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = \frac{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)} = 1$$

And third, because μ^+ is a measure function, and hence does not take negative values, $\mathrm{Ver}(\phi) \geq 0$, and because $[\![\phi]\!] \subseteq \prod_{k=1}^{w} \Pi(R^k)$, $\mathrm{Ver}(\phi) \leq 1$.

As for Axiom 2, first note that by the definition of $[\![\cdot]\!]$ and elementary set theory, it holds for all $\phi, \psi \in \mathcal{L}$ that (i) $[\![\phi \wedge \psi]\!] \cup [\![\phi \wedge \bar{\psi}]\!] = ([\![\phi]\!] \cap [\![\psi]\!]) \cup ([\![\phi]\!] \cap [\![\bar{\psi}]\!]) = [\![\phi]\!] \cap ([\![\psi]\!] \cup [\![\bar{\psi}]\!]) = [\![\phi]\!] \cap \prod_{k=1}^{w} \Pi(R^k) = [\![\phi]\!]$. In the same way, it holds that (ii) $[\![\phi \wedge \psi]\!] \cup [\![\bar{\phi} \wedge \psi]\!] = [\![\psi]\!]$. Hence, we have (iii) $[\![\phi \vee \psi]\!] = [\![\phi]\!] \cup [\![\psi]\!] = ([\![\phi \wedge \psi]\!] \cup [\![\phi \wedge \bar{\psi}]\!]) \cup ([\![\phi \wedge \psi]\!] \cup [\![\bar{\phi} \wedge \psi]\!]) = [\![\bar{\phi} \wedge \psi]\!] \cup [\![\phi \wedge \bar{\psi}]\!] \cup [\![\bar{\phi} \wedge \psi]\!]$. Further note that (iv) $[\![\phi \wedge \psi]\!] \cap [\![\bar{\phi} \wedge \bar{\psi}]\!] = [\![\phi]\!] \cap ([\![\psi]\!] \cap [\![\bar{\psi}]\!]) = \varnothing$, and that similarly (v) $[\![\phi \wedge \psi]\!] \cap [\![\bar{\phi} \wedge \psi]\!] = \varnothing$ and (vi) $[\![\bar{\phi} \wedge \psi]\!] \cap [\![\phi \wedge \bar{\psi}]\!] = \varnothing$. By the definition of Ver, it holds that

$$(1)\ \mathrm{Ver}(\phi) + \mathrm{Ver}(\psi) - \mathrm{Ver}(\phi \wedge \psi) = \frac{\mu+([\![\phi]\!]) + \mu+([\![\psi]\!]) - \mu+([\![\phi \wedge \psi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)}$$

© Douven and Decock 2017

By (i) and (ii), the right-hand side of (1) is equal to

$$(2)\quad \frac{\mu^+([\![\phi \wedge \psi]\!] \cup [\![\phi \wedge \bar{\psi}]\!]) + \mu^+([\![\phi \wedge \psi]\!] \cup [\![\bar{\phi} \wedge \psi]\!]) - \mu^+([\![\phi \wedge \psi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)}$$

By (iv) and (v) and the fact that μ^+ is a measure function, (2) is equal to

$$(3)\quad \frac{2\mu^+([\![\phi \wedge \psi]\!]) + \mu^+([\![\phi \wedge \bar{\psi}]\!]) + \mu^+([\![\bar{\phi} \wedge \psi]\!]) - \mu^+([\![\phi \wedge \psi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)},$$

which, by algebra, reduces to

$$(4)\quad \frac{\mu^+([\![\phi \wedge \psi]\!]) + \mu^+([\![\phi \wedge \bar{\psi}]\!]) + \mu^+([\![\bar{\phi} \wedge \psi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)}$$

And by (iv), (v) and (vi), and again the fact that μ^+ is a measure function, (4) is equal to

$$(5)\quad \frac{\mu^+([\![\phi \wedge \psi]\!] \cup [\![\phi \wedge \bar{\psi}]\!] \cup [\![\bar{\phi} \wedge \psi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)},$$

which, by (iii), is equal to

$$(6)\quad \frac{\mu^+([\![\phi \vee \psi]\!])}{\mu^+\left(\prod_{k=1}^{w} \Pi(R^k)\right)}$$

Finally, again by the definition of Ver, (6) is equal to $\mathrm{Ver}(\phi \vee \psi)$. ■

References

Abasnezhad, Ali and Otávio Bueno (eds.) 2017: *On the Sorites Paradox*. Berlin: Springer.

Adams, Ernest W. 1975: *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: Reidel.

Akiba, Ken 2004: 'Vagueness in the World'. *Noûs*, 38, pp. 407–44.

© Douven and Decock 2017

Berlin, Brent and Kay Paul 1969: *Basic Color Terms: Their Universality and Evolution.* Berkeley: University of California Press. Reprinted Stanford, CA: CSLI Publications, 1999.

Borg, Ingwer, Patrick J. F. Groenen, and Patrick Mair 2013: *Applied Multidimensional Scaling.* Heidelberg: Springer.

Churchland, Paul M. 2012: *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals.* Cambridge, MA: MIT Press.

Clark, Austen 1993: *Sensory Qualities.* Oxford: Clarendon Press.

Cook, Roy 2002: 'Vagueness and Mathematical Precision'. *Mind*, 111, pp. 225–47.

Decock, Lieven, Richard Dietz, and Igor Douven 2013: 'Modelling Comparative Concepts in Conceptual Spaces'. In Motomura, Butler and Bekki 2013, pp. 69–86.

Decock, Lieven and Igor Douven 2014: 'What Is Graded Membership?' *Noûs*, 48, pp. 653–82.

—— 2017: 'Two Recent Degree-Theoretic Approaches to the Sorites Paradox'. Forthcoming in Abasnezhad and Bueno 2017.

Dietz, Richard 2011: 'The Paradox of Vagueness'. In Horsten and Pettigrew 2011, pp. 128–79.

—— 2013: 'Comparative Concepts'. *Synthese*, 190, pp. 139–70.

Dietz, Richard and Sebastiano Moruzzi (eds.) 2010: *Cuts and Clouds: Vaguenesss, Its Nature and Its Logic.* Oxford: Oxford University Press.

Douven, Igor 2013: 'The Epistemology of Conditionals'. *Oxford Studies in Epistemology*, 4, pp. 3–33.

—— 2016a: *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches.* Cambridge: Cambridge University Press.

—— 2016b: 'Vagueness, Graded Membership, and Conceptual Spaces'. *Cognition*, 151, pp. 80–95.

Douven, Igor, Lieven Decock, Richard Dietz, and Paul Égré 2013: 'Vagueness: A Conceptual Spaces Approach'. *Journal of Philosophical Logic*, 42, pp. 137–60.

Douven, Igor and Sara Verbrugge 2013: 'The Probabilities of Conditionals Revisited'. *Cognitive Science*, 37, pp. 711–30.

Douven, Igor, Sylvia Wenmackers, Yasmina Jraissati, and Lieven Decock 2017: 'Measuring Graded Membership: The Case of Color'. *Cognitive Science*, (Forthcoming).

Earman, John 1992: *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* Cambridge, MA: MIT Press.

 © Douven and Decock 2017

Edgington, Dorothy 1992: 'Validity, Uncertainty and Vagueness'. *Analysis*, 52, pp. 193–204.
——1995: 'On Conditionals'. *Mind*, 104, pp. 235–329.
——1997a: 'Vagueness by Degrees'. In Keefe and Smith 1997, pp. 294–316.
——1997b: 'Commentary'. In Woods 1997, pp. 95–137.
Eells, Ellery, and Brian Skyrms (eds.) 1994: *Probability and Conditionals: Belief Revision and Rational Decision*. Cambridge: Cambridge University Press.
Fairchild, Mark D. 1998: *Color Appearance Models*. Reading, MA: Addison-Wesley.
Fara, Delia Graff 2000: 'Shifting Sands: An Interest-Relative Theory of Vagueness'. *Philosophical Topics*, 28, pp. 45–81.
——2010: 'Scope Confusions and Unsatisfiable Disjuncts: Two Problems for Supervaluationism'. In Dietz and Moruzzi 2010, pp. 373–82.
Fine, Kit 1975: 'Vagueness, Truth and Logic'. *Synthese*, 30, pp. 265–300.
Gärdenfors, Peter 2000: *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.
——2014: *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press.
Gärdenfors, Peter and Frank Zenker 2011: 'Using Conceptual Spaces to Model the Dynamics of Empirical Theories'. In Olsson and Enqvist 2011, pp. 137–53.
——2013: 'Theory Change as Dimensional Change: Conceptual Spaces Applied to the Dynamics of Empirical Theories'. *Synthese*, 190, pp. 1039–58.
Hájek, Alan 1989: 'Probabilities of Conditionals—Revisited'. *Journal of Philosophical Logic*, 18, pp. 423–8.
——1994: 'Triviality on the Cheap?' In Eells and Skyrms 1994, pp. 113–40.
Hájek, Petr 1998: *Metamathematics of Fuzzy Logic*. Berlin: Springer.
Hampton, James A. 2007: 'Typicality, Graded Membership, and Vagueness'. *Cognitive Science*, 31, 355–84.
Harper, William L. and Cliff A. Hooker (eds.) 1976: *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Volume I: Foundations and Philosophy of Epistemic Applications of Probability Theory*. Dordrecht: Reidel.
Horsten, Leon and Richard Pettigrew (eds.) 2011: *The Continuum Companion to Philosophical Logic*. London: Continuum.

 © Douven and Decock 2017

Jackson, Frank 1987: *Conditionals*. Oxford: Blackwell.
Kamp, Hans 1975: 'Two Theories about Adjectives'. In Keenan 1975, pp. 123–55.
Kamp, Hans and Barbara Partee 1995: 'Prototype Theory and Compositionality'. *Cognition*, 57, pp. 129–91.
Keefe, Rosanna 2000: *Theories of Vagueness*. Cambridge: Cambridge University Press.
——2012: 'Modelling Vagueness: What Can We Ignore?' *Philosophical Studies*, 161, pp. 453–70.
Keefe, Rosanna and Peter Smith (eds.) 1997: *Vagueness: A Reader*. Cambridge, MA: MIT Press.
Keenan, Edward L. (ed.) 1975: *Formal Semantics for Natural Languages*. Cambridge: Cambridge University Press.
Kleene, Stephen C. 1952: *Introduction to Metamathematics*. Amsterdam: North-Holland.
Krzyżanowska, Karolina, Sylvia Wenmackers, and Igor Douven 2013: 'Inferential Conditionals and Evidentiality'. *Journal of Logic, Language and Information*, 22, pp. 315–34.
——2014: 'Rethinking Gibbard's Riverboat Argument'. *Studia Logica*, 102, pp. 771–92.
Kyburg, Henry E., Jr. 1961: *Probability and the Logic of Rational Belief*. Middletown, CT: Wesleyan University Press.
Lewis, David K. 1970: 'General Semantics'. *Synthese*, 22, pp. 18–67.
——1976: 'Probabilities of Conditionals and Conditional Probabilities'. *Philosophical Review*, 85, 297–315.
Lycan, William 2001: *Real Conditionals*. Oxford: Oxford University Press.
Lyons, John 1995: *Linguistic Semantics: An Introduction*. Cambridge: Cambridge University Press.
MacFarlane, John 2010: 'Fuzzy Epistemicism'. In Dietz and Moruzzi 2010, pp. 438–63.
Motomura, Yoichi, Alastair Butler, and Daisuke Bekki (eds.) 2013: *New Frontiers in Artificial Intelligence*. Heidelberg: Springer
Nosofsky, Robert M. 1986: 'Attention, Similarity, and the Identification-Categorization Relationship'. *Journal of Experimental Psychology: General*, 115, pp. 39–57.
Okabe, Atsuyuki, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu 2000: *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd edn. New York: Wiley.
Olsson, Erik J. and Sebastian Enqvist (eds.) 2011: *Belief Revision Meets Philosophy of Science*. Berlin: Springer.

 © Douven and Decock 2017

Ramsey, Frank P. 1929: 'General Propositions and Causality'. In Ramsey 1990, pp. 145–63.
——1990: *Philosophical Papers*. Edited by D. H. Mellor. Cambridge: Cambridge University Press.
Sainsbury, R. M. 1990: 'Concepts Without Boundaries'. Reprinted in Keefe and Smith 1997, pp. 251–64.
——1991: 'Is There Higher-Order Vagueness?' *Philosophical Quarterly*, 41, pp. 167–82.
Smith, Nicholas J. J. 2008: *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.
——2012: 'Measuring and Modeling Truth'. *American Philosophical Quarterly*, 49, pp. 345–56.
Sorensen, Roy 1988: *Blindspots*. Oxford: Clarendon Press.
Stalnaker, Robert 1976: 'Letter to van Fraassen'. In Harper and Hooker 1976, pp. 302–6.
Temple, George F. J. 1971: *The Structure of Lebesgue Integration Theory*. Oxford: Clarendon Press.
van Deemter, Kees 2010: *Not Exactly: In Praise of Vagueness*. Oxford: Oxford University Press.
van Fraassen, Bas C. 1976: 'Probabilities of Conditionals'. In Harper and Hooker 1976, pp. 261–300.
Williamson, Timothy 1994: *Vagueness*. London: Routledge.
Woods, Michael 1997: *Conditionals*. Edited by David Wiggins, with a commentary by Dorothy Edgington. Oxford: Clarendon Press.
Zadeh, Lofti A. 1965: 'Fuzzy Sets'. *Information and Control*, 8, pp. 338–53.

© Douven and Decock 2017

Deflationism, Arithmetic, and the Argument from Conservativeness

Daniel Waxman
New York University
danielwaxman@gmail.com

Many philosophers believe that a deflationist theory of truth must conservatively extend any base theory to which it is added (roughly: talking about truth shouldn't allow us to establish any new claims about subject-matters not involving truth). But when applied to arithmetic, it's argued, the imposition of a conservativeness requirement leads to a serious objection to deflationism: for the Gödel sentence for Peano Arithmetic (PA) is not a theorem of PA, but becomes one when PA is extended by adding plausible principles governing truth. This paper argues that no such objection succeeds. The issue turns on how we understand the notion of logical consequence implicit in any conservativeness requirement, and whether we possess a categorical conception of the natural numbers (i.e. whether we can rule out so-called "non-standard models"). I offer a disjunctive response: if we possess a categorical conception of arithmetic, then deflationists have principled reason to accept a rich notion of logical consequence according to which the Gödel sentence follows from PA. But if we do not, then the reasons for requiring the derivation of the Gödel sentence lapse, and deflationists are free to accept a conservativeness requirement stated proof-theoretically. Either way, deflationism is in the clear.

1. Introduction

Deflationism about truth is one of the major contemporary answers to the question of the nature of truth. The best way of briefly introducing the view is to say that it stands in opposition to the well-known 'traditional' accounts of truth, which attempt to give a substantive analysis of what the truth of a sentence or proposition consists in.[1] According to the correspondence theory, for instance, it is whether a sentence corresponds to the world that determines its truth; and according to the coherence theory, it is its inclusion in some (perhaps privileged) coherent set of sentences. Unlike these theories and others like them, deflationism rejects the idea that truth plays any serious explanatory role in philosophy or elsewhere.

[1] As far as possible, I want to ignore the issue of what the bearers of truth are; in particular, I will stay neutral between those who embrace propositions and those who prefer sentences (however these are typed). For purposes of brevity and readibility, I will use 'sentences', with the proviso hereby made that I stand officially agnostic on this issue.

Mind, Vol. 126 . 502 . April 2017 **© Waxman 2017**
doi:10.1093/mind/fzv182 Advance Access publication 19 January 2017

In this paper, I want to discuss and assess a particular line of attack that has been pressed against deflationists in recent years, originating in articles by Stewart Shapiro (1998) and Jeffrey Ketland (1999). These authors contend that the distinctive commitments of deflationism force a recognition that a formal theory of truth must be *conservative* over the truth-free theory to which they are added ('the base theory'). Loosely put, this requirement states that any sentence in the base language that does not already follow from the base theory also does not follow from the truth theory. (Note the slippery term 'follows from', which is ambiguous between (at least) two distinct notions of logical consequence. This will be discussed very soon below.) But, they think, this gets the deflationist into trouble in the case of arithmetic: the Gödel sentence (G) for Peano Arithmetic (PA) is not a theorem of PA itself, but becomes a theorem once PA is extended by adding certain natural principles governing truth. The resulting theory of arithmetical truth is strong enough to prove the Gödel sentence, and is therefore non-conservative over arithmetic itself. And this, it is argued, poses a serious problem for deflationists, because (according to its critics) deflationism is committed to truth theories being conservative over their bases.

In broad outline, the deflationist has three potential avenues of response. The most basic choice-point concerns the question whether deflationists are committed to conservativeness in some sense or another. What I will call *rejectionist* responses simply answer no. *Compatibilist* responses are those that maintain that deflationism is both committed to and compatible with a conservativeness constraint. Such views have a further question to answer: in the sense that deflationists are committed to conservativeness, what is the relevant notion of logical consequence employed in its formulation? *Proof-theoretic* compatibilists answer that it is proof-theoretic or deductive consequence. Responses of this sort contend that a deflationist can in good conscience accept a conservative theory of arithmetical truth, either by denying the need for G to be a theorem of the truth theory or by sketching a different—as it were, a truth-free—route to its derivation. *Semantic* compatibilists accept a conservativeness requirement, but insist that it is to be understood in terms of a semantic notion of logical consequence, according to which G *is* a genuine logical consequence of the axioms of arithmetic.[2] Responses of this

[2] To forestall potential misunderstanding, 'semantic' here is used to mean, roughly, 'model-theoretic', as opposed to 'having to do with meaning'.

 © Waxman 2017

sort necessarily involve going beyond the proof-theoretic resources of first-order predicate logic: as we know, by the completeness theorem for first-order arithmetic, semantic and proof-theoretic consequence coincide for such languages.

All of the present defenders of deflationism in the literature of which I am aware are either rejectionists or proof-theoretic compatibilists. By contrast, semantic compatibilism has had no explicit defenders to date.[3] Although an attempt to force the deflationist into the acceptance of a rich notion of logical consequence was at the forefront of Shapiro's original paper, it has been for the most part ignored by those in the deflationist camp in favour of a more direct rebuttal. Volker Halbach goes so far as to call it 'more a trap than a way out' (Halbach 2001, p. 170). But a moment's consideration will show that this possibility is of more relevance than has been acknowledged. In a debate which hinges on, among other things, the nature of arithmetic, the question naturally arises: how is arithmetic to be understood? As I see the issue, the crux is whether we possess what is often called a categorical conception of the natural numbers—an understanding of arithmetic sufficiently rich to rule out the so-called non-standard models that arise for the first-order axiomatization—or whether arithmetic is understood in more axiomatic terms, as (merely) a formal theory in first-order logic. The issue of whether we have a categorical conception of the natural numbers, and if so, how, is a huge one, and I will not be able to discuss it in depth here. Nevertheless, I shall argue that this distinction is crucial for finding our way sensibly through the conservativeness argument. Once we see things clearly, the deflationist can respond disjunctively as follows, without taking a stand on which of the disjuncts obtains. If we have a categorical conception of arithmetic, then this fact alone provides us with reason to countenance a stronger-than-proof-theoretic notion of logical consequence, and the deflationist in particular is put at no disadvantage by doing so. But if we do not, then the reasons for requiring the derivation of the Gödel sentence lapse. Either way, the deflationist is in the clear.

By the end, I hope to have accomplished three main goals. The first is to have clarified the debate, and in particular to have dispelled some of the confusion surrounding the way in which arithmetic and various notions of logical consequence figure in it. The second is to have

[3] Horwich (1990) is perhaps committed to such a view. See footnote 7 for further discussion.

 © Waxman 2017

explained why deflationists need not fear notions of semantic consequence that are stronger than proof-theoretic consequence—or, at least, no more than anyone else. The third is to have shown, through the viability of the disjunctive strategy mentioned above, that the form of conservativeness that deflationists should endorse is sensitive to the underlying theory of arithmetic. Far from being an unwelcome consequence, this is as it should be: for deflationists, truth is explanatorily inert, and so it should come as no surprise that one's conception of arithmetic will have a major impact on how one's resulting theory of arithmetical truth will turn out.

The plan of the rest of the paper is as follows. §2 briefly looks at the diversity of views that go under the heading of deflationism, and draws out some preliminary constraints that such views impose on formal theories of truth. §3 introduces and summarizes some technical material concerning formal theories of arithmetical truth, considers some of the different ways one might expand *PA* via the addition of a truth predicate, and states results about the conservativeness of these theories over *PA* itself. After these preliminaries, §4 presents in detail the non-conservativeness objection to deflationism. §5 discusses several of the most influential deflationist responses, and (while finding much of worth) argues that none is fully satisfying. I then go on in §6 to argue for the disjunctive response I advocate on behalf of the deflationist. Finally, §7 contains some concluding remarks about the most plausible forms of deflationism going forward.

2. Deflationism and formal theories of truth

The immediate difficulty faced by any attempt to say something informative about deflationism, or assess any argument presented against it, is that it comes in many varieties. As I introduced it earlier, deflationism is the rejection of the thesis that there is any philosophically heavyweight theory of truth. But this basic idea can be, and has been, fleshed out in several different ways. Here I distinguish four strands of deflationism and attempt to draw out some (preliminary) implications that they have for formal theories of truth.

The first strand of deflationism is *metaphysical*: it claims that truth is a metaphysically thin, or deflated, or disunified property; or more radically still, that there is no single property that true sentences have in common. Although this kind of terminology is perhaps suggestive, it is very difficult to see how to develop the idea, owing to the

 © Waxman 2017

murkiness of the notions involved. To the extent that the metaphysical formulation is appealed to by the proponents of the argument from conservativeness against deflationism, it is used in an unanalysed and wholly intuitive way.

Secondly, there is a *conceptual* or *semantic* strand of deflationism, concerning the *concept* of truth or the *meaning* of the predicate 'is true'.[4] The core claim of this variety is that the conceptual or semantic content of truth is *thin*: it is, in some sense to be elaborated, given by nothing more than the T-schema:

T-schema $T(\ulcorner \phi \urcorner) \leftrightarrow \phi$

Different varieties of conceptual and semantic deflationism carry out this elaboration in different ways. Here are two of the most influential examples. On Paul Horwich's *minimalist* view (Horwich 1990), it is constitutive of an English speaker's understanding of 'is true' that he is disposed to accept the (infinitely many) English instances of the T-schema.[5] The concept of truth is then defined derivatively as the common constituent of belief states expressed in uses of the word by those who understand it. For Horwich, that is the sense in which all there is to truth is given by the T-schema.[6]

As a result of his minimalist commitment that our understanding of truth is underwritten by the instances of the T-schema, Horwich faces what has come to be known as the generality problem. The problem is this: the bare instances of the T-schema seem to be too weak for deflationist purposes. In particular, they do not entail principles such as the law of excluded middle, $\forall p(T(p) \vee T(\neg p))$, and the compositional laws, for instance, for conjunction:

$$\forall p \forall q(T(p \wedge q) \leftrightarrow (T(p) \wedge T(q)))$$

that, intuitively, a theory of truth should be able to prove. Consequently, there is a real question as to how minimalists are able to earn themselves the right to endorse such generalizations.[7]

[4] Although one might have reason to distinguish these two claims— the semantic and the conceptual—for our purposes, it would introduce irrelevant complications. We can take the claim indifferently to concern the *content* of truth, without worrying here whether content is to be understood linguistically or mentally.

[5] Strictly speaking, Horwich's view is stated in terms of propositions rather than sentences, but for our purposes nothing turns on this subtlety.

[6] I use corner-brackets as a naming device. See footnote 12 for more on notation, and §3.2 for a brief discussion of the implications of various choices available here.

[7] See Gupta (1993) for an early critique of deflationism on these grounds (although, as Hartry Field pointed out to me, the relevant results were already known to Tarski), as well as §3 below.

 © Waxman 2017

Hartry Field's position is different. For Field, the centrality of the T-schema is given by the fact that for any speaker who understands ϕ, there is a cognitive equivalence between ϕ and $T(\ulcorner\phi\urcorner)$, with the notion of cognitive equivalence cashed out in broadly inferentialist terms. However, to overcome the generality problem, Field wants to go further than merely accepting the instances of the T-schema. He explores one way of doing this by working directly with a *schematic* version of the T-schema. The idea is to exploit our facility with schematic reasoning directly—as it were, with schematic variables entering into the object language rather than the metalanguage. In short, schematic variables are introduced into the language, and rules of inference are introduced that allow one to substitute sentences for schematic variables and to infer $\forall x(Sent(x) \rightarrow P(x))$ from the schema $P(\ulcorner\phi\urcorner)$.[8] This allows one to derive generalizations such as the compositional laws from the schematic version of the T-schema. I will have a little more to say about this schematic approach in §6.

A third thesis of deflationism is that the positive role played by truth is solely an *expressive* or *generalizing* one. This claim is often combined with the other varieties of deflationism above, but is also occasionally taken to constitute the primary claim of deflationism itself. In particular, it is held that the primary purpose of the notion of truth is to allow us to indirectly endorse sentences to which we are able to refer obliquely but whose content we are unable or unwilling to specify fully (for example, 'What the Pope said yesterday is true' or 'The Gödel sentence for *PA* is true'), and to endorse whole classes of sentences, including infinite sets, simultaneously (for example, 'Everything the Pope says is true' or 'All of the theorems of *PA* are true'). The idea that the notion of truth serves a primarily expressive or generalizing role is a natural fit with the conceptual and semantic varieties of deflationism. For it provides an elegant explanation of

---

This problem becomes yet more pronounced when it is claimed by minimalists such as Horwich that the primary *purpose* of (the notion of) truth is to establish generalizations (see the third deflationist thesis below); for in that case, the minimalist's truth predicate is unable adequately to carry out the very task for which it is designed. Horwich is, of course, aware of the issue, and to settle it proposes a rule of inference analogous to the ω-rule (see §5) which will allow the troublesome generalizations to be obtained. However, this commits him to the acceptance of a notion of semantic consequence stronger than proof-theoretic. Although Horwich never, to my knowledge, discusses the argument from conservativeness, this suggests that he may be friendly to what I have called a semantic compatibilist view, for once the ω-rule is added as a rule of inference, *G* becomes a consequence of the axioms of *PA*.

[8] Details of the schematic approach can be found in the appendix to Field (2001, ch. 4) and in Field (2006).

 © Waxman 2017

the—otherwise seemingly recalcitrant—fact that truth-talk is a well-entrenched, and indeed communicatively essential, part of natural language. If the predicate 'is true' is governed by principles as apparently trivial as the T-schema, one might reasonably wonder why we see fit to employ it at all, let alone to the extent we do; and the expressive or generalizing use of 'true' provides an explanation of these facts that is at least plausible on its face.

Fourthly and finally, there is a thesis concerning the *explanatory* role of truth. On this understanding, deflationism is the claim that truth does not play a central explanatory role in our best theories of the world, including, but not limited to, the philosophies of mind and language, intentional psychology and semantics. As Field has emphasized, there is a crucial caveat: if truth figures essentially in an explanation *due only to its expressive or generalizing role*, that is perfectly acceptable in the eyes of most deflationists.

Deflationism is thus opposed to views that hold that truth plays a more robust explanatory role in our understanding of the world. Barry Loewer gives a helpful list of some reasonably prevalent philosophical doctrines that would, if correct, conflict with deflationism so construed:

> [That] truth is a substantive value and goal for belief and assertion, that understanding a language and grasping thoughts consists in knowledge of truth conditions (and other semantic properties and relations), that the difference between factual and non-factual discourse is explicated in terms of semantic notions, and that semantic properties (and relations) of thoughts and other intentional mental events and states enter into rationalizing and causal explanations of actions. (Loewer 2005, p. 59)

Needless to say, this list is not exhaustive, and indeed the argument from conservativeness can be viewed as attempting to add an additional entry: that an inflationary notion of truth allows us, furthermore, to explain the fact stated by the Gödel sentence for *PA*.

In what follows, the form of deflationism I am interested in is the view constituted by the conjunction of a version of the semantic or conceptual claim, the expressive or generalizing claim, and the explanatory claim. In particular, I take both Field's disquotationalism and Horwich's minimalism to be paradigmatic deflationary views in the sense I will consider.

Let us now ask: what requirements does deflationism, so understood, place upon a formal theory of truth? I take the foregoing discussion to motivate at least the following three conditions, each

 © Waxman 2017

corresponding to one strand of the view. It must (1) include or allow the derivation of every instance of the T-schema,[9] and (2) allow the expression and derivation of generalizations, including generalizations involving the notion of truth, all while (3) ensuring that truth plays no essential explanatory role that does not derive from its expressive or generalizing capacities. Conditions (2) and (3) are perhaps less than fully clear; but they will receive further discussion as we proceed.

3. Some technical preliminaries

3.1 Conservativeness

For the sake of definiteness, I take a theory to be simply a set of sentences understood as its axioms—not, as the notion is sometimes used, a set of sentences closed under logical consequence. Let $\mathcal{T}$ be a theory stated in a language $\mathcal{L}$ (the 'base language'). Now consider an extended language $\mathcal{L}^+$, and let $\mathcal{T}^+$ be a theory extending $\mathcal{T}$ and stated in $\mathcal{L}^+$. The general notion of a conservative extension is defined as follows:

> $\mathcal{T}^+$ is a *conservative extension* of $\mathcal{T}$ if, for any sentence ϕ in the base language $\mathcal{L}$, if ϕ follows from $\mathcal{T}^+$ then ϕ follows from $\mathcal{T}$ alone.

The definition above contains the ambiguous term 'follows from', indicating that a species of logical consequence is in play. So we can disambiguate between (at least) the following two species of conservativeness:

> $\mathcal{T}^+$ is a *deductively (or proof-theoretic) conservative extension* of $\mathcal{T}$ if, for any sentence ϕ in the base language $\mathcal{L}$, if ϕ is deducible (in a given proof system) from $\mathcal{T}^+$ then ϕ is deducible from $\mathcal{T}$ alone.
>
> $\mathcal{T}^+$ is a *semantically conservative extension* of $\mathcal{T}$ if, for any sentence ϕ in the base language $\mathcal{L}$, if ϕ is true in all models of $\mathcal{T}^+$ then ϕ is true in all models of $\mathcal{T}$ alone.

Owing to the soundness and completeness of (all of the natural) deductive systems for first-order logic, these two notions are coextensive in the case where such a deductive system is used and where $\mathcal{T}$ is a first-order theory. But as we will later be considering notions of semantic consequence in which they come apart, the distinction

[9] At least (to avoid confrontation with the Liar and other semantic paradoxes) for sentences stated in the truth-free fragment of the language.

 © Waxman 2017

between the two species of conservativeness is well worth keeping in mind in what follows.

3.2 Some ways of adding a truth predicate to PA

Let $\mathcal{L}$ be the usual language of arithmetic, and let *PA* be the theory consisting of the usual first-order axioms of Peano arithmetic (stated, of course, in $\mathcal{L}$). *PA* is an infinitely axiomatized theory, in particular containing every instance of the induction schema for formulas ϕ containing vocabulary in $\mathcal{L}$:

Induction schema $(\phi(0) \wedge \forall x(\phi(y) \rightarrow \phi(Sx))) \rightarrow \forall x\phi(x)$

The question we are interested in is: how might we extend *PA* to obtain a theory of arithmetical truth? As we shall see, there are several ways in which this can be done.

Syntactically, the natural approach is to move to a new language $\mathcal{L}^+ = \mathcal{L} \cup T$ whose new one-place predicate T is to be understood as truth. To predicate truth of sentences we require a means of naming them, and we do this in the usual way by assuming a Gödel numbering so that $\ulcorner\phi\urcorner$ denotes the Gödel number of ϕ and E_n is the expression whose Gödel number is n. Then $T(n)$ is to be understood as 'E_n is true'.

All of the theories we examine are extensions of *PA*, that is, every axiom of *PA* will be one of their axioms. Furthermore are all *materially adequate*, in Tarski's sense, that is, they entail each instance of the T-schema:

T-schema $T(\ulcorner\phi\urcorner) \leftrightarrow \phi$

where ϕ is a sentence stated in $\mathcal{L}$.[10]

Here, then, are four theories to consider:[11]

(1) PA_D is the theory whose axioms are those of *PA* together with every instance of the T-schema for sentences stated in $\mathcal{L}$. This is perhaps the most direct way of obtaining a materially adequate truth theory for *PA*—in essence, by adding the minimally required *T*-sentences by brute force.

(2) PA_D^+ is the theory whose axioms are those of PA_D together with the instances of the induction schema required to extend it in the following way. Whereas *PA* includes an

[10] Note the restriction to sentences stated in $\mathcal{L}$: we cannot, on pain of serious engagement with the paradoxes, require that our theories satisfy the T-schema where ϕ is allowed to contain the truth predicate.

[11] The names of the different theories are adopted from Shapiro (2002).

 © Waxman 2017

instance of the induction schema for every sentence stated in the truth-free language $\mathcal{L}$, PA_D^+ includes also an instance for every sentence *stated in the extended language* $\mathcal{L} \cup T$. In other words, PA_D^+ extends induction to deal with cases including the truth predicate.

Philosophical remark: PA_D^+ and PA_D might be thought of as formal implementations of Horwichian minimalism about truth (see §2): someone who believes that 'all there is to say' about truth is given by the instances of the T-schema will plausibly adopt one of these two theories (the choice between them depending on whether they wish to extend arithmetical induction to sentences containing T).

(3) PA_T is obtained by adding to PA Tarskian compositional clauses for the logical connectives. There are several equivalent ways in which this can be done (some of which first define satisfaction, and then define truth in terms of that). Here is a representative set of such axioms:[12]

- $\forall s \forall t(T(s = t) \leftrightarrow s^\circ = t^\circ)$
- $\forall p(T(\neg p) \leftrightarrow \neg T(p))$
- $\forall p \forall q(T(p \wedge q) \leftrightarrow T(p) \wedge T(q))$
- $\forall x \forall p(T(\forall x p) \leftrightarrow \forall t\, T(p[t/x]))$

(4) PA_T^+ is the same as PA_T except that it also extends induction to include sentences containing T.

[12] An explanation of the notation (which allows for greater perspicuity at the cost of suppressing the details of the coding). Quantification into sentence position is indicated by variables p and q, and is to be understood as shorthand for quantificiation restricted to a formula, $Sent(x)$, satisfied by a number n if and only if n is the Gödel number of a sentence of $\mathcal{L}$. (Which formula it is will depend on fine-grained—and for our purposes, irrelevant—details of the Gödel numbering.) Similarly for quantification into term position (using variables s and t). I use the symbol $^\circ$ as an abbreviation for the evaluation function for terms. As a convenience, I write quantifiers and connectives within the scope of the T predicate, although to be precise we would need to talk of functions from Gödel numbers to Gödel numbers representing the effect of applying those quantifiers and connectives. So, for instance, when written out fully, the axiom for negation would become $\forall x(Sent(x) \rightarrow (T(\dot{\neg} x) \leftrightarrow \neg T(x))$, where $\dot{\neg}$ represents the function that takes the Gödel number (x) of a sentence to the Gödel number of its negation. Finally, $\forall t\, T(p[t/x])$ expresses the truth of the sentence resulting from substituting the variable x with a term t within the sentence coded by p. See Halbach (2011, p. 33) for more details.

 © Waxman 2017

Let us turn now to a examination of the strength of the four theories discussed above. PA_D and PA_D^+ are both conservative extensions of *PA*.[13] However, both have difficulty proving generalizations concerning truth. For instance, although each theory can prove, *for each sentence* ϕ, $T(\ulcorner\phi\urcorner) \vee T(\ulcorner\neg\phi\urcorner)$, neither can prove the universal generalization of this fact:

$$\forall p(T(p) \vee T(\neg p))$$

This is a consequence of the compactness of first-order logic: for the generalization to follow, it would have to follow from finitely many of the axioms; but it clearly does not follow from any finite set of *T*-sentences.

PA_T is also a conservative extension of *PA*.[14] Unlike PA_D and PA_D^+, PA_T is very good at proving generalizations; for instance, it is able to prove bivalence, as well as analogous principles for other connectives.[15]

Not all expressible generalizations fare so well, however. Consider the following principle:

T-Bew $\forall p(Bew_{PA}(p) \rightarrow T(p))$—all the theorems of *PA* are true.

Although all the *instances* of T-Bew are provable in all four theories we are examining, the general principle itself is not provable in PA_D and PA_D^+ (by a simple compactness argument) and in PA_T. By contrast, PA_T^+ *does* prove T-Bew, and as a result is able to formalize the following argument, resulting in its proof-theoretic non-conservativeness over *PA*.[16] To see this, note that (since the right-hand side is an instance of T-Bew):

$$PA_T^+ \vdash Bew_{PA}(\ulcorner 0 = 1\urcorner) \rightarrow T(\ulcorner 0 = 1\urcorner)$$

[13] For PA_D and PA_D^+, here is a quick sketch of the proof. Suppose we can derive ϕ. The derivation of ϕ consists of a finite number of sentences; so there must be an upper bound on the complexity of these sentences—suppose it is Σ_n. Now, it is possible within *PA* to define a partial truth predicate T_n for Σ_n-sentences. We can thereby obtain a *PA* proof of ϕ by replacing the instances of the T-schema used in the original derivation with instances of the T_n-schema (together with explicit proofs of the T_n-sentences appealed to). For a detailed proof see Halbach (2011).

[14] See Halbach (2011, p. 100) for a proof-theoretic argument to this effect.

[15] Thanks to Jeffrey Ketland for pointing out a mistaken claim about PA_T in an earlier draft.

[16] This argument is in all essentials the 'semantic argument' discussed in more depth in §4.2.

 © Waxman 2017

From the T-schema (whose instances are derivable in PA^+_T) we also have:

$$PA^+_T \vdash T(\ulcorner 0 = 1 \urcorner) \leftrightarrow 0 = 1$$

and, combining the two, we get:

$$PA^+_T \vdash Bew_{PA}(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1$$

But of course *PA* proves that $0 \neq 1$, as does its extension PA^+_T. Contraposing and applying *modus ponens*, we obtain:

$$PA^+_T \vdash \neg Bew_{PA}(\ulcorner 0 = 1 \urcorner)$$

But $\neg Bew_{PA}(\ulcorner 0 = 1 \urcorner)$ is just Con_{PA}, the statement that *PA* is consistent, and we know by Gödel's Second Incompleteness Theorem that (if *PA* is consistent, which I henceforth assume without comment) $PA \nvdash Con_{PA}$. Therefore, PA^+_T is a non-conservative extension of *PA*.

This completes the brief review of the technical situation. The following table succinctly summarizes the crucial results.

	Truth axioms	Extends induction?	Proves T-Bew?	Conservative over *PA*?
PA_D	*T*-instances	No	No	Yes
PA^+_D	*T*-instances	Yes	No	Yes
PA_T	Compositional	No	No	Yes
PA^+_T	Compositional	Yes	Yes	No

4. The argument from conservativeness

The argument from conservativeness can be viewed as operating in three steps. The first step is that deflationism imposes the constraint that truth theories be conservative, in at least some sense of the notion. The second step contends that any adequate theory of arithmetical truth will be *proof-theoretically* non-conservative. The third step argues that for deflationists, the relevant notion of consequence figuring in the conservativeness constraint is proof-theoretic, and that deflationists cannot avail themselves of a stronger, semantic notion. In the taxonomy introduced in §1, those deflationists who resist the first step are rejectionsts; those who accept the first step and resist the second are proof-theoretic compatibilists; and those who accept the

 © Waxman 2017

first and second steps while resisting the third are semantic compatibilists. I will discuss each step in turn.

4.1 Deflationism requires conservativeness

Here is the core of Shapiro's argument for a conservativeness requirement. (Note that nothing in this passage suggests that conservativeness need be understood in proof-theoretic terms.)

> I submit that in one form or another, conservativeness is essential to deflationism. Suppose, for example, that Karl correctly holds a theory B in a language that cannot express truth. He adds a truth predicate to the language and extends B to a theory B' using only axioms essential to truth. Assume that B' is not conservative over B. Then there is a sentence b in the original language (so that b does not contain the truth predicate) such that b is a consequence of B' but not a consequence of B. That is, it is logically possible for the axioms of B to be true and yet b false, but it is not logically possible for the axioms of B' to be true and b false. This undermines the central deflationist theme that truth is insubstantial. Before Karl moved to B', $\neg\phi$ was possible. The move from B to B' added semantic content sufficient to rule out the falsity of b. But by hypothesis, all that was added in B' were principles essential to truth. Thus, those principles have substantial semantic content. (Shapiro 1998, p. 497)

Ketland, along similar lines:

> One might suggest that these [conservativeness results] illustrate a kind of 'analyticity' or 'contentlessness' that deflationary theories of truth exhibit. Adding them 'adds nothing'. Indeed, it is these metalogical properties that are closely connected to the idea that the deflationary truth theories illustrate the 'redundancy' or 'non-substantiality' of truth. Indeed, one might go further: if truth is non-substantial as deflationists claim then the theory of truth should be conservative. Roughly: non-substantiality $\equiv$ conservativeness. (Ketland 1999, p. 79)

A preliminary point to make is that such arguments cannot be right as they stand. Halbach (2001) has pointed out that any adequate truth theory—including the pure minimalist theory (consisting of the addition of the instances of the T-schema)— will not be conservative over pure logic. For take some logical truth, ϕ. Any adequate truth theory will prove the instances of the T-schema for ϕ and $\neg\phi$, and therefore will prove both $T(\ulcorner\phi\urcorner)$ and $\neg T(\ulcorner\neg\phi\urcorner)$. It follows that $\exists x\exists y(x \neq y)$, a conclusion that does not follow from pure logic alone.

However, an appropriate and non-ad hoc modification of the conservativeness claim is easy to come by. The essence of the problem raised by Halbach is that a theory of truth is committed to the

 © Waxman 2017

existence of *bearers* of truth. As mentioned earlier, it is controversial what exactly truth-bearers are. But that there must be truth-bearers of one kind or another is not seriously up for debate; and whatever they are, a formal theory of truth must include names for them. This requirement is satisfied by the inclusion of a theory of syntax, containing a name for each potential truth-bearer statable in the base language. Typically, syntax is simply arithmetized: a scheme of Gödel numbering is implemented, setting up a correspondence between natural numbers and potential truth-bearers. Setting up a Gödel numbering means that the theory of syntax is taken just to be *PA*; but this is not the only possible way to proceed. Equally well, one might choose to name potential truth-bearers with strings of symbols, in which case the theory of syntax would be a theory of strings and concatenation. However this is done, the obvious amendment of the conservativeness requirement is that the truth theory be conservative over the base theory *plus the theory of syntax*. The conservativeness claim must, and can reasonably, be amended in this way.[17]

4.2 Adequacy requires proof-theoretic non-conservativeness

Why do Ketland and Shapiro believe that any adequate truth theory will be (proof-theoretically) non-conservative? The answer concerns our ability to recognize the truth of the Gödel sentence for arithmetic. Here are the opening lines of an essay by Dummett:

> By Gödel's theorem there exists, for any intuitively correct formal system for elementary arithmetic, a statement [*G*] expressible in the system but not provable in it, which not only is true but can be recognised by us to be true … (Dummett 1963, p. 186)

Dummett here adds an epistemic claim that is not strictly contained in the formal mathematical result: that *G can be recognized by us to be true*. A natural question to ask is: how does this recognition proceed? As orthodoxy has it, it goes via an argument using the concept of arithmetical truth. Here again is Dummett on that argument:

> By hypothesis the axioms of the system are intuitively recognised as being true, and the rules of inference of the system as being correct … Hence we may establish by an inductive argument on the lengths of formal proofs that each proof in the system has a true conclusion, and by another inductive argument on the number of logical constants in a statement that no statement is both true and false; concluding from this that the system is

[17] Shapiro (2002) accepts this amendment to his original statement of the conservativeness requirement. See Heck (forthcoming) for a helpful discussion of alternative theories of syntax.

 © Waxman 2017

> consistent [a statement which is provably equivalent in the system to *G*]. (Dummett 1963, p. 195)

Let us call this the *semantic argument for G*.[18] Shapiro puts it as follows:

> [O]nce our subject has taken on the truth predicate, and he notices that all the axioms of [*PA*] are true and that the rules of inference preserve truth, he concludes that every theorem of [*PA*] is true. He also knows (from [the T-schema]) that '0 = 1' is not true, and so '0 = 1' is not a theorem of [*PA*]. So our subject concludes that [*PA*] is consistent ([Con_{PA}]) and that [*G*] is true. *The defect of [any weaker theory] as a theory of arithmetic truth is that it cannot reproduce this simple, informal reasoning.* (Shapiro 1998, p. 499, emphasis added)

Ketland (1991, p. 91) similarly places heavy weight on the ability of PA^+_T to *prove* *G*, and thereby allow us to come to recognize its truth. For otherwise—without a proof from a theory of arithmetical truth—he thinks it is difficult, if not impossible, to explain why *G* is true.

It is worth distinguishing two desiderata for a deflationist theory of arithmetical truth that are not clearly separated by Shapiro and Ketland, one weak and one strong. The weak desideratum is that the deflationist must be able to derive the (truth of the) Gödel sentence by giving *some* proof or other from principles concerning arithmetic and truth that are acceptable from a deflationist perspective. The strong desideratum is that it must be possible to formalize the semantic argument itself in the deflationist's theory of arithmetical truth. As we will see, Tennant's response (discussed in §5) is an example of a deflationist response that satisfies the weak desideratum but not the strong one.

As was mentioned in §3.2, PA_D and PA^+_D fail to yield a great many generalizations which intuitively it seems an adequate theory of truth should yield. However, PA_T *does* yield many such generalizations, in particular the ones arising via the interaction of the connectives with the truth predicate. The compositional principles it includes are responsible for this additional power (although, as we have seen, they are not powerful enough to yield T-Bew). It might then be thought that PA_T is an attractive middle ground for the deflationist to inhabit—it is strong enough to account for many, though not all, generalizations involving truth, and yet weak enough

[18] In effect, the proof of *G* in §3 is just the formalization of this argument in a sufficiently powerful theory of truth—it is precisely this result that leads to the non-conservativeness of PA^+_T over *PA*.

 © Waxman 2017

to still be proof-theoretically conservative over *PA*. It is worth mentioning Shapiro's response to this move: he wants to deny such an option to the deflationist, for he thinks that the position is an unstable one:

> [W]hether one is a deflationist or not, there is no good reason to demur from the extension of the induction scheme to the new language. There is no reason to demur from $[PA^{+}_{T}]$. Informally the induction principle is that for any well-defined property (or predicate), if it holds of 0 and is closed under the successor function, then it holds of all natural numbers. It does not matter if the property can be characterized in the original, first-order theory. (Shapiro 1998, p. 500)

In support of this claim, he cites Dummett: 'It is part of the concept of natural number, as we now understand it, that induction with respect to any well-defined property is a ground for asserting all natural numbers to have that property' (Dummett 1994, p. 337). As we shall later see, the idea that induction must be extended to any property will prove to be a watershed between two different understandings of arithmetic.

4.3 The move to stronger notions of logical consequence

Assuming the success of the argument so far, what room for manoeuvre is left for the deflationist? The only escape route is the acceptance of a notion of logical consequence (or a notion of what is 'implicit in' a theory) that goes beyond its proof-theoretic consequences—what I earlier called semantic compatibilism. For if the deflationist can appeal to such a notion, the possibility arises of arguing that the deductive consequences of PA^{+}_{T} in the language of arithmetic—including *G*—were 'all along', so to speak, consequences of our best truth-free theory of arithmetic. The sting would then be taken from the conservativeness argument: for although the theory of arithmetical truth would not be proof-theoretically conservative over arithmetic, it would nevertheless be semantically conservative for the relevant species of semantic consequence, thereby satisfying the intuition behind the conservativeness requirement.

Shapiro (1998), however, thinks that this option is a 'thin reed' for the deflationist to take. His discussion focuses primarily on two difficulties. Firstly, all of the available options for adding logical resources to obtain a strengthened notion of semantic consequence appear to have substantial logical and mathematical content. Take second-order semantic consequence under the standard semantics, for instance.

 © Waxman 2017

Many mathematical structures—for example, the natural numbers and the real numbers—possess categorical second-order axiomatizations; even the universe of sets (about as rich a structure as is imaginable) is characterized 'quasi-categorically', in the sense that any two models of second-order Zermelo-Fraenkel set theory with the Axiom of Choice are either isomorphic or one is isomorphic to an initial segment of the other. As a result, truth in all of these structures can be defined in terms of second-order consequence. The second-order approach in particular faces the additional worry that second-order logic involves an expansion in one's ontology, with some (most famously, Quine) going so far as to attribute to it all of the ontology of set theory. Secondly, any notion of consequence strong enough to deliver the set of arithmetical truths is bound, by Gödel's results, to be non-effective. This poses an epistemological problem: if there is no effective procedure by which we can recognize the validity of an inference, it is not clear how we can be said to truly grasp the notion of consequence at stake. At the very least, Shapiro thinks, these are obstacles which render this approach extremely unattractive.

5. Deflationist responses to the argument

5.1 Can the conservativeness requirement be denied?

Is there any reasonable scope for denying that a deflationist theory of truth must be conservative? Shapiro and Ketland both emphasize that the conservativeness requirement draws upon what I have termed the metaphysical understanding of deflationism: that 'truth is insubstantial', or some similar claim. This is not indisputable: the formulation of the metaphysical construal is somewhat murky, and in particular, very little is said in support of the crucial transition from 'truth is insubstantial' to 'truth theories must be conservative'. Nevertheless, the transition has considerable intuitive force, for it seems extremely uncomfortable to maintain that truth is an insubstantial or non-robust property if the addition of truth principles leads one to rule out what were previously considered to be live possibilities concerning a (truth-free) subject matter. Perhaps the best way of understanding the transition is as a proposed explication: the informal notion of metaphysical insubstantiality is to be (possibly partially) explicated in terms of the formal criterion of conservativeness. It is striking, and a mark in favour of the plausibility of this understanding,

 © Waxman 2017

that the conservativeness requirement has attracted considerable support among deflationists themselves.

It might be thought that an analogous argument for the conservativeness requirement could be made, appealing not to the deflationist's committment to the metaphysical insubstantiality of truth, but rather to its explanatory inertness. On reflection, I am doubtful that such an argument succeeds.[19]

Suppose B is our base theory and B^+ results from B by adding appropriate truth principles. Let *explanatory conservativeness* then be the claim:

> For any sentence ϕ in the base language, if ϕ is explained by B^+, then ϕ is explained by B.

An argument could be run against the deflationist if it could be shown that (i) PA^+_T explains G, (ii) PA fails to explain G, and (iii) the deflationist is committed to explanatory conservativeness.

I think that (i) is arguably true. Although not all proofs are explanatory, the semantic argument for G very plausibly is.[20] So, since the semantic argument can be run in PA^+_T, we have every reason to believe that PA^+_T thereby explains G.

I am much more doubtful of (ii). The best reason I can think of for endorsing (ii) appeals to the claim that proof-theoretic derivability is a necessary condition for mathematical or logical explanation. But—to anticipate issues that arise in §5—if we countenance a notion of logical consequence richer than proof-theoretic derivability, this condition will appear unmotivated.

Nevertheless, even if (i) and (ii) are granted for the sake of argument, the Gödel sentence nevertheless fails to cause a problem for the deflationist, because (iii) is false. Recall from §2 that the deflationist finds utility in the notion of truth precisely because it plays a generalizing and expressive role. At best, then, explanatory conservativeness is a constraint to which the deflationist is committed only when dealing with explanations *in which the role of truth does not derive solely from its expressive or generalizing capacities.* As Field puts it, 'any use of "true" in explanations which derives solely from its role as a device of generalization should be perfectly acceptable' (1999, p. 537). But the role played by truth in the semantic argument is precisely a

[19] Thanks to an anonymous reviewer for pressing me to clarify an earlier version of this argument.

[20] See Shapiro (1998, p. 505) for additional argument to this effect.

 © Waxman 2017

generalizing one: as Shapiro acknowledges, the 'centrepiece of the explanation' is the ability of truth to 'make the generalization' that 'every theorem of *PA* is true' (1998, p. 506). So I am sceptical whether considerations of explanatory inertness can be marshalled in service of an argument against the deflationist here.

5.2 Interlude: Two conceptions of arithmetic

Before we continue discussing how the deflationist can respond to the argument from conservativeness, we need to get clear on a more basic issue first. What do we mean when we talk about 'arithmetic'? There is an ambiguity here, underscored by two different conceptions of the subject matter. The first conception is broadly model-theoretic in nature: for lack of a better term, let us call it the *categorical* conception of arithmetic. It holds that arithmetic is a subject about a particular mathematical structure—the natural numbers, N, the elements of which are obtained by beginning with o and interating the successor operation finitely many times. By contrast, the other conception is broadly proof-theoretic in nature: let us call it the *axiomatic* conception of arithmetic. It holds that our best understanding of arithmetic consists in (and is exhausted by) the proof-theoretic consequences of a particular set of axioms, namely, first-order *PA*.[21] As is well known, these two conceptions of arithmetic come apart. First-order *PA* has so-called non-standard models, models which are not isomorphic to N.[22] For the categorical theorist, such models are to be ruled out as *unintended interpretations* of the axioms, in contrast to N, which is often called the intended interpretation. On the categorical conception, arithmetic is not fully adequately captured by its axiomatization in first-order *PA*, or indeed any first-order axiomatization, for any such first-order theory will admit non-standard models. If the axiomatic conception is viewed through a model-theoretic lens, arithmetic will

[21] For arguments that first-order *PA* occupies an "intrinsic, conceptually well-defined region of arithmetical truth", see Isaacson (1996). Despite singling out first-order *PA* in this way, Isaacson is not an adherent of the axiomatic conception, for he holds that higher-order resources are implicated in our full understanding of the structure of the natural numbers.

[22] One way to see this uses the Löwenheim-Skolem theorem: any first-order theory with infinite models has models of every infinite cardinality, so there are models of *PA* of cardinality $> \aleph_0$. Another way is to expand the language of arithmetic by adding a new constant symbol, c, and consider the theory obtained by adding to *PA* the set of sentences $\{s_n : n \in N\}$, where s_n is $c > \bar{n}$. Every finite subset of this theory has a model—if k is the largest number such that s_k is in our subset, interpret c as $k + 1$. So, by the compactness theorem, that theory has a model. But although this model satisfies *PA*, it also contains a non-standard number which is, so to speak, 'infinitely distant' from o.

 © Waxman 2017

appear to have a diversity of admissible models, none of which has any claim to being privileged as 'intended' over and above the others.

It is a fundamental question in the philosophy of mathematics which of the conceptions is more defensible.[23] Although there is a natural tendency amongst mathematicians to take something like the categorical conception for granted, for a variety of reasons some philosophers have been sceptical of our ability to 'intend' a unique interpretation. Those who have tried to sustain a categorical conception have often sought to do so by endowing their logic with additional resources of sufficient strength to identify a single model (or, at least, a class of isomorphic models). There are many ways in which this can be done, if one is only willing to countenance the required resources. Here are a few possibilities.

Firstly, one might appeal to second-order quantification with the 'standard' semantics.[24] The second-order Peano axioms differ from the first-order axioms only in that they replace the (infinitely many instances of the) induction schema with a single induction *axiom*, $\forall X(X(\mathrm{o}) \wedge \forall y(X(y) \rightarrow X(Sy))) \rightarrow \forall x X(x)$. Non-standard models then do not arise, for second-order *PA* is a categorical theory—any two models are isomorphic. The existence of non-standard models can be explained by noting that the first-order induction schema allows induction only over predicates definable in the language of arithmetic; and since there are subsets of N that are not so definable, the induction schema is therefore bound to 'miss' genuine instances of induction, and thereby 'over-generate' models of arithmetic. Alternatively, if one has scruples about appealing to full second-order quantification—perhaps for reasons of ontological commitment, or epistemic worries about the intractibility of the second-order consequence relation—there are other options. One option is adopting the logic L_{Q_0}, involving the addition of an additional cardinality quantifier whose interpretation is stipulated to be 'there exist at most finitely many __'. The consequence relation for L_{Q_0} is non-effective, but not as rich as full second-order consequence (for instance, L_{Q_0} does not characterize the real numbers up to isomorphism) (see Shapiro 2001). In the same vicinity is ω-logic, which licenses the inference from (the infinitely many premisses)

[23] Not only in the philosophy of arithmetic: similar questions arise in the philosophy of set theory.

[24] The standard semantics take the second-order quantifiers to range over all sub-collections of the domain, as opposed to Henkin semantics, which takes them to range over a distinct (and possibly truncated) 'second-order' domain.

 © Waxman 2017

$\phi(0), \phi(1), \phi(2)\ldots$ to $\forall x \phi(x)$.[25] A further alternative—which will perhaps be attractive to deflationists who emphasize the use of the notion of truth in expressing infinite sets of sentences—is the adoption of an infinitary logic such as $L_{\omega_1, \omega}$, which allows the construction of infinite conjunctions and disjunctions. As with second-order logic, these alternatives characterize the natural numbers uniquely up to isomorphism, and so non-standard models do not arise.

Accepting any of these logical resources will commit one to accepting a semantic consequence relation according to which *G* and all other arithmetical truths are genuine semantic consequences of arithmetic. This is one reason for some philosophers' unease with the categorical conception: on one plausible interpretation of Gödel's theorems, the set of arithmetical truths is not recursively enumerable, and so it is questioned whether we can have any grasp of a non-effective consequence relation that generates all such truths.

5.3 Reflection principles to the rescue?

Tennant (2002) takes what I have called a proof-theoretic compatibilist view: he accepts that deflationism is committed to the proof-theoretic conservativeness of an arithmetical truth theory over *PA*, but thinks that this will pose no threat to deflationism. In fact, he thinks that the deflationist position involves the claim that truth is not a genuine property, and that this motivates deflationists to refuse to extend induction to expressions containing the truth predicate, the intuitive idea being that we are required to extend induction only to genuine properties. So, in particular, Tennant denies that a deflationist theory of truth need have the (non-conservativeness-inducing) implication T-Bew, the generalization to the effect that all the theorems of *PA* are true. But since he accepts that we do have reason to recognize the truth of *G*, he offers, as he must, an alternative explanation of how this is possible. To cut a long story short, his favoured approach employs the adoption of (a rule of inference corresponding to) the following reflection principle for primitive recursive formulae:

PR-Reflection $\forall x Bew_{PA}(\ulcorner \phi(x) \urcorner) \rightarrow \forall x \phi(x)$ (where ϕ is primitive recursive)

PR-Reflection is (provably within *PA*) equivalent to *G* and Con_{PA}, which leads Tennant to conclude that it is 'just what is needed in order to be able to formalize faithfully the reasoning in the [semantic

[25] For more on ω-logic and the ω-rule, see Isaacson (1992).

 © Waxman 2017

argument]' (2002, p. 573).[26] For the Tennant-style deflationist, the appeal of explaining our justification for recognizing the truth of G via this principle is clear: the move allows him to accept a conservative *truth theory*, PA_D, while, so to speak, pinning the blame for non-conservativeness on the reflection principle—a principle which (it is claimed) does not need to appeal to considerations of truth for its motivation.[27]

Ketland criticizes Tennant on this point:

> *If* Tarski's theory of truth [i.e. PA^+_T] provides at least *one* way of 'recognizing the truth of Gödel sentences', then this fact alone contradicts deflationism (for a deflationary theory of truth should be conservative). The entirely different assumption, which I did not make, that this is 'the only way' is irrelevant. (Ketland 2005, p. 83, emphasis in original)

Ketland seems here to say that the mere *existence* of a non-conservative theory of arithmetical truth that proves G poses a threat to deflationism. But this line of thought only goes through on the premiss—which is the crux of the current dispute, and which I take it that Tennant will deny— that the relevant theory of arithmetical truth (in this case, PA^+_T) is properly understood as *deflationary*. The alternative, reflection-principle-based explanation of the truth of G is precisely an attempt to undermine such a premiss by arguing that the deflationist need not be committed to PA^+_T; so this attempted short way with the Tennant-style deflationist begs the question and fails to pose a real threat. As far as I can tell, there is no simple outright refutation of Tennant's strategy. Rather, its attractiveness rests on the answers to two questions. Firstly, to what extent can PR-Reflection be plausibly motivated independently of the truth-theoretic considerations it is supposed to sidestep? And secondly, how problematic is the insistence that the truth theory fails to prove T-Bew?

I do not think the first issue poses much of a problem, although I have little to say in its support over and above what Tennant already

[26] Tennant makes much of the fact that PA^+_T proves, not only Con_{PA}, but $Con_{PA+ConPA}$, $Con_{PA+ConPA+ConConPA}$, and so on, while adding (the rule form of) PR-Reflection only enables the proof of Con_{PA}. So, he contends, the truth theorist is guilty of using a stronger tool than is required. However, I do not find this line of thought convincing in support of Tennant's proposal. For that proposal relies on the restriction of the reflection principle to only primitive recursive sentences, and no independent motivation for such a restriction is given. If *unrestricted* reflection is accepted and iterated, then Tennant's theory too will be substantially stronger than required. See Feferman (1962, §5).

[27] Since Tennant makes clear that he himself is not a deflationist, there are, strictly speaking, no Tennant-style deflationists.

 © Waxman 2017

says. In brief, the idea is that our acceptance of reflection principles is licensed by our acceptance of the theory itself; such principles are simply formalizations of our commitment to accept whatever follows from the theory. It may be possible to argue that acceptance of PR-Reflection is tacitly underwritten by some truth-theoretic considerations, but I do not see how such an argument would run.

The second issue is, in my view, far more damaging. As has been emphasized, a central thread running through deflationism emphasizes the *expressive utility* of the notion of truth in allowing us to state and prove generalizations. On this point, Field writes that 'it is quite uncontroversial that the notion of truth can be used to make generalizations, and that these generalizations can be important to giving rise to commitments not involving the notion of truth' (1999, p. 536). But now, the sentence $\forall p(Bew_{PA}(p) \rightarrow T(p))$—T-Bew—seems like precisely such a generalization. It is not as if the Tennant-style deflationist will want to *deny* that all the theorems of *PA* are true, in the sense of wanting a truth theory from which its negation follows; for even on the most austere deflationist theories—those based on adding the bare instances of the T-schema—it is provable *of each individual theorem* that the relevant instance is true.[28] Such an austere stance carries with it a sense of refusing to live up to one's commitments; and what is worse, these commitments are precisely generalizations of the kind that the notion of truth is supposed to help the deflationist formulate and derive. Consequently, Tennant's approach fails to satisfy what in §4.2 I called the strong desideratum for a theory of truth: although he is arguably able to explain the truth of the Gödel sentence, he is not able to carry out the semantic argument for *G*—intuitively, a cogent argument, and one that a theory of arithmetical truth ought to be able to capture. So, at the very least, Tennant's strategy incurs substantial costs, and motivates the exploration of alternatives.

[28] Why does PR-Reflection fail to establish T-Bew even in the presence of the T-schema? Two reasons. Firstly, note that PR-Reflection is restricted to primitive recursive sentences; T-Bew, by contrast, is a claim about *all* sentences, not just those that are primitive recursive (or of the form $\forall x \phi(x)$, where ϕ is primitive recursive). Secondly, even if PA_D + PR-Reflection yielded all the instances of T-Bew, the further step of collecting them up into the universal generalization—T-Bew itself—would remain to be established, and as mentioned in §3.2, the bare instances of the T-schema are bad at establishing generalizations concerning truth. For a rigorous proof, see Smorynski (1977).

 © Waxman 2017

5.4 First-order deflationism

I want now to examine Jody Azzouni's (1999) defence of what he calls the first-order deflationist—essentially, a deflationist who accepts what we have called the axiomatic conception of arithmetic, and whose understanding is consequently insufficient to rule out the eligibility of non-standard models.[29] Like Tennant, Azzouni is a proof-theoretic compatibilist: he wants to abstain from extending the induction schema to sentences containing the truth predicate and accept proof-theoretic conservativeness, which rules out the possibility of proving T-Bew. He comes down on the side of accepting what we have called PA_T. To fend off attacks based on the inability of this theory to prove certain generalizations, he states that such generalizations are not 'essential to truth'. I am not sure how convincing this line of thought is, because of the difficulty in fleshing out what is really 'essential to truth' and what is merely extraneous. For instance, Azzouni accepts the need for a theory of truth to prove generalizations such as the compositional axiom for conjunction, and it is difficult to see a principled reason why *that* kind of generalization is 'essential to truth' while others like T-Bew are not.[30]

But Azzouni's grounds for refusing to extend induction are more interesting. He notes that the intuitive considerations that Shapiro adduces in support of extending induction apply only to the *standard model*—that is, they are cogent only if one has the categorical conception of arithmetic in mind. Recall Shapiro's claim that 'the induction principle is that for any well-defined property (or predicate), if it holds of o and is closed under the successor function, then it holds of all natural numbers. It does not matter if the property can be characterized in the original, first-order theory'. But these considerations hardly apply to those—like proponents of the axiomatic view—who profess to acknowledge only first-order-definable properties in inductions. Proponents of such a view will indeed regard the extension of induction as ruling out perfectly legitimate models of *PA*. For this reason, Azzouni thinks, the first-order deflationist can escape the burden of Shapiro's strictures; and he seems to imply that only by going first-order can the deflationist so escape.

Although this response is inventive, it nevertheless has the undesirable property of committing the deflationist to a particular conception

[29] Perhaps a better name would be a 'proof-theoretic' deflationist.

[30] On similar grounds, Shapiro (2002) thinks that the inability of the first-order deflationist to prove T-Bew is in itself a *reductio* of that position.

 © Waxman 2017

of arithmetic: the axiomatic conception. For although this may in the final analysis be the only defensible conception, it is far from obvious that this is so. Many will contend that it is simply a datum that we *do* have an intended interpretation of theories of arithmetic—*the natural numbers*—and that we *can* make sense of the distinction between standard and non-standard models. To underscore just one of the many implications of Azzouni's view, it follows that G is not an arithmetical truth altogether, for (even if *PA* is consistent) there are some non-standard models in which G is false! If deflationism requires an axiomatic conception of arithmetic, then we are faced with a judgement of relative plausibility: what is more plausible, that we are able to grasp the standard model of arithmetic or that deflationism is correct? At best, deflationism is on highly controversial ground, and at worst it is landed with an impoverished and inadequate understanding of arithmetic.

Despite these critical remarks, I think that Azzouni's observation—that the extension of induction to formulae containing the truth predicate implictly appeals to considerations motivated only by the standard model—is insightful, and underscores a fundamental ambiguity in the debate about how arithmetic is to be understood. And while I do not think that being forced into adopting the axiomatic view is necessarily an attractive solution for the deflationist, attending to the ambiguity will allow the deflationist to respond satisfactorily to the non-conservativeness objection. In that connection, let us turn to the response mooted in Field (1999).

5.5 Diagnosing the source of non-conservativeness

Like Azzouni and Tennant, Field seems to embrace the proof-theoretic conservativeness requirement. But unlike both authors, Field advocates the extension of induction to include instances involving the truth predicate. As he puts it:

> [P]eople who commit themselves to the arithmetic induction schema mean to be committing themselves not only to its instances in the language they have but to all instances in any legitimate expansion, including one with a truth predicate. (Field 1999, p. 539)

As mentioned, one advantage of adopting a theory like PA^{+}_{T} that extends induction is that it enables the proof of generalizations such as T-Bew. This is, of course, not news to Field, who emphasizes that it is of a piece with a deflationist understanding of the notion of truth:

 © Waxman 2017

> [T]he main point of having the notion of truth, many deflationists say, is that it allows us to make fertile generalizations we could not otherwise make; where by a fertile generalization I mean one that has an impact on claims not involving the notion of truth. (Field 1999, p. 533)

This combination of views would appear to put Field directly in the crosshairs of the non-conservativeness argument. His response is distinctive and suggestive. In essence, he claims that the extended induction axioms derive from *facts about the natural numbers*, rather than *facts about truth*. As he puts it, what the extended induction instances involving the truth predicate depend on is 'a fact about the natural numbers, namely, that they are linearly ordered with each element having only finitely many predecessors' (Field 1999, p. 538). In this regard, extending induction to include sentences involving the truth predicate is no different from extending induction by including any other new predicate defined in one's language: in both cases the extension is licensed by facts about the nature of the natural numbers, and in neither case is there any reason to expect the extension to behave conservatively. The strategy is, in short, to diagnose non-conservativeness as stemming from the 'arithmetic' part of 'arithmetical truth'. We might put it this way: the move from PA to PA_T is justified on truth-theoretic grounds, while the move from PA_T to PA_T^+—the move that introduces non-conservativeness—is justified on number-theoretic grounds; so it is arithmetic, and not truth, that is to blame for non-conservativeness. It is important to note that Field has in mind a categorical conception of arithmetic: the number-theoretic grounds that Field appeals to in extending induction are ones that assume the falsity of the axiomatic conception of arithmetic. The giveaway is the appeal to 'finitely many predecessors', a notion that cannot fully be captured using only first-order proof-theoretic resources.[31] Field assumes that our understanding of the natural numbers transcends their axiomatization in first-order PA, and that it is on the basis of this understanding that we are justified in extending induction.[32]

As attractive as it is, I believe that there is a serious tension in Field's view. In §1, I distinguished between proof-theoretic compatibilists—

[31] Naturally, it can be captured in first-order set theory, but the problem of non-standard interpretations for set theory is just as acute.

[32] This is not quite an accurate representation of Field's position, since he accepts a theory in which the T-schema is used schematically in the object language; but his theory is close enough to PA_T^+ that this sloppiness is justified on expositional grounds.

 © Waxman 2017

those deflationists who accept a proof-theoretic conservativeness constraint for truth theories—semantic compatibilists—those who accept a conservativeness constraint formulated in terms of some stronger-than-proof-theoretic notion of semantic consequence—and rejectionists—those who deny any kind of conservativeness requirement. The question is, which type does Field's response fall under? On the one hand, he seems to accept proof-theoretic conservativeness as a constraint: as he puts it, 'there is no need to disagree with Shapiro when he says conservativeness is essential to deflationism'. But the problem with this reading is that Field does in fact accept a non-conservative theory of arithmetical truth. Of course, he tries to deflect the blame away from truth and on to arithmetic, but this deflection is an *excuse* for non-conservativeness, not a *denial* of it. So, it seems, Field's response is best interpreted as being outright rejectionist. For the reasons discussed in §5.1, I think that rejectionism is an unattractive way for the deflationist to go.

That being said, I think that Field's discussion contains an important insight, namely, that the additional strength of the theory obtained by extending induction can be attributed to arithmetic, and not truth. Unlike Field, and as I will argue in the next section, I think this insight can be used—to the extent that we have reason to adopt (as Field does) a categorical conception of arithmetic—to motivate not a rejectionist view but a semantic compatibilist one.

6. The disjunctive strategy

I believe that the deflationist is in a position to mount a compelling response to the argument from non-conservativeness and to Shapiro's charge that a deflationist must move to a non-effective notion of consequence. The response stands neutral on the question of whether we have most reason to adopt a categorical conception of arithmetic or an axiomatic conception, and proceeds disjunctively depending on how this question is ultimately best resolved. The crucial idea is that on either disjunct, the view is going to end up being a compatibilist one, for both disjuncts are going to end up accepting *some* conservativeness requirement or other. But the notion of logical consequence in which the relevant conservativeness requirement is stated—proof-theoretic or semantic—will depend on the operative understanding of arithmetic. Let me explain each disjunct in turn.

 © Waxman 2017

Suppose, on the one hand, that an axiomatic conception of arithmetic is most defensible. On this conception, our understanding of the subject matter of arithmetic is constituted by our understanding of a canonical first-order axiomatization of arithmetic (let us continue to assume that this is given by the axioms of first-order *PA*). I take this to mean that, as far as sentences in the language of arithmetic are concerned, the axiomatic theorist is committed to accepting all and only those that are deductive consequences of *PA*. Our question, then, is what truth theory a deflationist who adheres to such a conception should adopt. My answer is that such a deflationist has a principled reason to accept PA_T (or even perhaps something weaker still), and in particular to demur from accepting PA^+_T. The reason for resisting the move to PA^+_T more or less falls out of the axiomatic characterization of arithmetic: extending induction to cover sentences containing T allows the derivation of G—a sentence that is (on this view) not licensed by the background understanding of arithmetic, since it is not derivable from the axioms.[33] Against the background of an axiomatic conception of arithmetic, then, the deflationist can and should accept a proof-theoretic conservativeness requirement with a clear conscience; for whichever of the theories of arithmetical truth weaker than PA^+_T is accepted, it will be proof-theoretically conservative over the relevant base.[34]

Suppose, on the other hand, we can be shown to be working with a categorical conception of arithmetic, one that cannot be captured in first-order terms. If that is the case, then we must somehow be in possession of resources that enable us to rule out non-standard models as being non-standard. However these resources are best characterized, they will induce a notion of arithmetical consequence that goes beyond merely what can be derived from *PA* in a given formal proof system. But if we think such a notion of consequence is in good standing and is implicit in our conception of arithmetic, then surely *it*,

[33] The situation will be one in which the axiomatic theorist is committed to accepting the disjunction $G \vee \neg G$—as an instance of the law of excluded middle, it clearly follows from *PA*—but not committed to accepting either of the disjuncts. In general, I take it that there is nothing incoherent about such a stance.

[34] What about the objection that *any* adequate theory of truth for a theory T will allow the proof of the reflection principle 'All theorems of T are true'? (An objection of this form is implicit in Ketland 1999, and was pressed upon me by Jeffrey Ketland in correspondence.) I do not deny that this is an attractive feature for a theory of truth to possess; nevertheless, I believe it presupposes the falsity of the axiomatic conception of arithmetic, for it requires the acceptance of truths in the language of arithmetic that do not follow (in the relevant sense) from its axioms.

 © Waxman 2017

and not proof-theoretic consequence, is what should figure in the relevant conservativeness requirement. And if conservativeness is understood in this way—in terms of an enriched notion of consequence—then PA^{+}_{T} will indeed be a conservative extension, for G (and in fact, for reasons of categoricity, all other true sentences in the language of arithmetic) will be a consequence of the (suitably enriched) base theory. In other words, I am recommending that, on this disjunct, the deflationist accept PA^{+}_{T} (or something stronger still) and embrace *semantic* compatibilism.[35] The proof-theoretic result on which the anti-deflationist argument rests will come to appear of dubious relevance, for—given a categorical conception of arithmetic—formal derivations will be seen as a legitimate but non-exhaustive way of drawing out the genuine consequences of the axioms. There is no reason for it to be at all objectionable that the addition of a theory of truth is non-conservative when this unduly weak notion of consequence is considered. In short, it is the *conception of arithmetic as categorical, as specifying a unique mathematical structure*, that is

[35] The reason for the qualification that the categorical deflationist might want to move to a truth theory *stronger* than PA^{+}_{T} is simply that whatever additional logical resources one appeals to in order to attain categoricity, one will presumably want one's eventual theory of truth to apply to sentences formulated in terms of them. (Thanks to an anonymous reviewer for raising this point and the need to address it.) And here, of course, PA^{+}_{T} is insufficient, for it is only a theory of truth for sentences in the *first*-order language of arithmetic, not for sentences in the newly enriched language of arithmetic (however this enrichment is carried out). Let me take the second-order formulation of arithmetic as a representative example and mention two ways in which a truth theory can be obtained. (Both of these extensions are more naturally formulated in terms of satisfaction rather than truth, but as mentioned in §3.2, this amounts only to a cosmetic difference). Firstly, one could work against a set-theoretic background and follow Shapiro (1991, p. 72) in generalizing the notion of a variable assignment to include values of second-order variables and adding the obvious satisfaction clauses (e.g. s satisfies $\forall X\phi$ iff, for every variable assignment s' that differs from s only (at most) in its assignment to X, s' satisfies ϕ). Truth can then be defined as usual as satisfaction by all variable assignments. Or—perhaps philosophically cleaner for a committed second-order theorist—one could follow Rayo and Uzquiano (1999) in employing second-order quantification and handling variable assignments by taking them to be the values of second-order variables obeying certain conditions.

To be sure, either one of these approaches will require logical or mathematical machinery beyond first-order *PA*: the Shapiro-style approach appeals to enough set theory to assign sets or sets of n-tuples of individuals as the values of second-order variables, and the Rayo and Uzquiano approach appeals to second-order quantification. But this should be neither surprising nor objectionable in the present dialectical context. It should not be surprising that a truth theory for second-order arithmetic will require resources going beyond those involved in giving a truth theory for first-order arithmetic. And appealing to such resources is legitimate, for extending one's truth theory in this way will still result in a conservative extension over *PA* (in the enriched sense of conservativeness that, I have argued, the deflationist on this disjunct of the response ought to adopt).

 © Waxman 2017

responsible for proof-theoretic non-conservativeness, not anything to do with truth.

Naturally, the deflationist taking this line needs to have something to say in reply to the reasons why Shapiro thinks that this option—that of adopting a strengthened notion of consequence—is a 'thin reed' for the deflationist to take. Recall that there were two complaints, the first concerning the substantive ontological character of the additional logical resources, the second concerning our ability to epistemically grasp the resources required to induce a non-effective consequence relation. I think the reply here ought to run basically as follows. If Shapiro's objections are cogent, then they are objections to the very possibility of a categorical conception of arithmetic. If the logical resources required for a categorical conception of arithmetic are ontologically problematic, or we are somehow epistemologically unable to grasp them, then that is a problem for *anyone* who maintains that arithmetic has an intended interpretation. But if that somewhat pessimistic diagnosis turns out to be correct, still no problem for deflationism has been disclosed: the deflationist can (like everyone else) coherently sidestep the worries by adopting an axiomatic conception of arithmetic. I claimed earlier that the cost of doing this was that it forecloses the possibility of a categorical conception of arithmetic; but if the option of a categorical conception is already foreclosed because of (what are by hypothesis) sound objections against the notion of consequence that it involves, then the first-order deflationist position begins to look more attractive.

Someone taking Shapiro's line might respond in the following way: it is not that the additional logical resources in question are problematic in themselves, but rather, they are problematic *specifically when pressed into service by the deflationist*. However, I see no reason to think that this is the case, and to my knowledge no such reason can be found in the literature. Here I will just give some prima facie arguments in support of this claim. Take the ontological–mathematical worry first. Deflationism is in the first instance a thesis about truth, not about the abstract ontology implicated by or the substantive mathematical character of certain logical resources. If there is a problem with the ontology or mathematical character of second-order logic (or whatever other option is adopted), then this is surely a problem for everyone, not just deflationists. Shapiro tentatively advances a contrary line:

> There are second-order categorical characterizations of just about every major mathematical structure, including the natural numbers, the real

 © Waxman 2017

numbers, the first inaccessible rank of the set-theoretic hierarchy, and beyond, well into the hierarchy of large cardinals. Therefore, truth for each of those theories can be reduced to second-order logical consequence ... [A] critic of deflationism might respond to the second-order manoeuvre by arguing that the deflationist is hiding the robustness of truth in the second-order consequence relation. If the consequence relation is itself robust, then the contemplated manoeuvre fails to show that truth is thin and has no nature. (Shapiro 1998, p. 510)

However, this line of thought begs the question against the categorical deflationist.[36] That truth in arithmetic and set theory can be defined in terms of second-order consequence is a firm technical result, not in dispute. But what *is* in dispute is whether truth in these theories is robust, that is, unacceptable to deflationists. To hold that the deflationist 'is hiding the robustness of truth' in the consequence relation *presupposes* that truth in the relevant domains is indeed robust—a presupposition that simply begs the question in the present context. What the critic of deflationism needs here are independent reasons that the deflationist in particular has commitments that block the adoption of the second-order consequence relation, and I am not aware that any such reasons can be given. Of course, this is not conclusive. But I believe the burden is on the anti-deflationist: *what specifically about deflationism* makes the mathematical character or ontological commitments of second-order consequence (or the alternatives mooted above) problematic? In what way does it build in a commitment to inflationary conceptions of truth?

The epistemological worry fares similarly. The question is again, why is the *deflationist* in a uniquely bad position in accounting for our grasp of non-effective consequence relations? It is true that such consequence relations are highly computationally intractable, and there are very pressing questions about how epistemic agents like us can grasp them (if indeed we can). I am not in possession of an account of how we are able to grasp such relations, so I cannot say for sure what the answer will look like. But if it is to pose a special problem to deflationism, it must presumably be argued that our grasp of such relations depends on a robust conception of truth; and it strikes me as difficult to construct an argument to this effect. How would conceiving of truth as correspondence to the facts, or as coherence, or pragmatically, or whatever, help here? None of these

[36] This is not a criticism of Shapiro, for it is clear from the text that he has reservations about the argument he presents.

 © Waxman 2017

conceptions of truth appear to allow us somehow to transcend the computational and epistemic limitations with which we are faced, or make it any easier to grasp a non-effective consequence relation; or at least, if they do, I cannot see how. I do not know how to refute the idea that the deflationist faces a special problem here, without providing a concrete proposal as to how the problem can be solved. But if the anti-deflationist wishes to take this line, surely he faces the burden of explaining how inflationary truth is able to provide a solution.

Let me briefly sum up the discussion of this section. The deflationist can adopt a conciliatory, disjunctive response to the argument from conservativeness, hedging his bets with respect to the question of how arithmetic ought to be understood. In advance of resolution of this question, there are two paths open to the deflationist: (i) to adopt a categorical conception of arithmetic, accept PA^{+}_{T} (or perhaps a stronger theory) as a theory of arithmetical truth, accept a strengthened notion of semantic consequence, and accept a conservativeness constraint formulated in terms of that notion; or (ii) to adopt an axiomatic conception of arithmetic, accept PA_T (or perhaps some other truth theory that is proof-theoretically conservative over *PA*), and accept a proof-theoretic conservativeness constraint. Which option is more attractive will depend on the tenability of categorical conceptions of arithmetic; but either way, the deflationist is in the clear.

7. Concluding remarks: Where does deflationism go from here?

Where does this all leave deflationism? As I have argued, the deflationist who believes that our understanding of arithmetic is sufficient to rule out non-standard models has independent reason to accept logical resources stronger than proof-theoretic logic. Return again to Field's proposal of employing schematic principles and reasoning as a means of deriving the compositional principles and solving the generality problem—this is as attractive a version of expanding one's logical resources as any. It has been charged that schematic reasoning of this form is tantamount to admitting a non-effective consequence relation, for *PA* (with induction formulated as an open-ended schema) is a categorical theory.[37] But there is a straightforward sense in which this non-effectiveness can be seen to derive from the

[37] The system is equivalent to a subsytem of second-order logic allowing the expression of Π^1_0 sentences. See Shapiro (2002) for discussion.

 © Waxman 2017

'indefinite extensibility' of the language (as it might be put): if we restrict the stock of predicates to a *fixed* set, the consequence relation becomes once again effective. This, I think, goes a very long way in assuaging the epistemological worries that Shapiro raises, for it is no longer as if we are faced with unrecognizable and unverifiable claims of the validity of certain inferences (as, arguably, say, full second-order logic would commit us to).

What about ontological worries, to the effect that accepting schematic reasoning commits us to recognizing an ontology of sets? The question of ontological commitment of schemata is even less resolved than the analogous question for full second-order logic (which is itself an issue of considerable controversy).[38] At the very least, it is not *clear* that employing schemata commits one to recognizing sets or anything comparably problematic, and there is a long tradition (including Quine himself) of using schemata precisely as a means of avoiding full second-order logic.

I do not mean to pin my colours to the mast of schematic reasoning here; but I do think that at the very least it provides a promising way for the deflationist to reconcile the desire for a categorical conception of the natural numbers with the need for an epistemically tractable notion of consequence.[39]

References

Azzouni, Jody 1999: 'Comments on Shapiro'. *Journal of Philosophy*, 10, pp. 541–4.

Barwise, John (ed.) 1977: *Handbook of Mathematical Logic*. Amsterdam: North-Holland.

Detlefsen, Michael (ed.) 1992: *Proof, Logic, and Formalization*. London: Routledge.

Dummett, Michael 1963: 'The Philosophical Significance of Gödel's Theorem'. Reprinted in Dummett 1978, pp. 186–214.

——1978: *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press.

[38] See McGee (1997) for an influential paper advocating the thesis that schemata avoid certain undesirable commitments of second-order logic; and see Pedersen and Rossberg (2010) for an opposing argument to the effect that they are on the same footing.

[39] Many thanks to Ralf Bader, Cian Dorr, Hartry Field, Yu Guo, Paul Horwich, Daniel Isaacson, Jeffrey Ketland, Penn Lawrence, Harvey Lederman, Noel Swanson, Jared Warren, Crispin Wright and Mike Zhao for helpful discussion and comments. Thanks are also due to two anonymous reviewers and the editor of *Mind* for valuable comments.

 © Waxman 2017

——1994: 'Reply to Crispin Wright'. In McGuinness and Oliveri 1994, pp. 329–38.
Feferman, Solomon 1962: 'Transfinite Recursive Progressions of Axiomatic Theories'. *Journal of Symbolic Logic*, 27, pp. 259–316.
Field, Hartry 1999: 'Deflating the Conservativeness Argument'. *Journal of Philosophy*, 10, pp. 533–40.
——2001: *Truth and the Absence of Fact*. Oxford: Clarendon Press.
——2006: 'Compositional Principles vs. Schematic Reasoning'. *Monist*, 89(1), pp. 9–27.
Gabbay, Dov, and Franz Guenthner (eds.) 2001: *Handbook of Philosophical Logic*, vol. 1. 2nd edn. Dordrecht: Kluwer.
Gupta, Anil 1993: 'A Critique of Deflationism'. *Philosophical Topics*, 1, pp. 57–81.
Halbach, Volker 2001: 'How Innocent is Deflationism?' *Synthese*, 126, pp. 167–94.
——2011: *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.
Hart, William (ed.) 1996: *The Philosophy of Mathematics*. Oxford: Oxford University Press.
Heck, Richard forthcoming: 'The Logical Strength of Compositional Principles'. *Notre Dame Journal of Formal Logic*.
Horwich, Paul 1990: *Truth*. Oxford: Oxford University Press.
Horsten, Leon, and Volker Halbach (eds.) 2002: *Principles of Truth*. Frankfurt: Hänsel-Hohenhausen.
Isaacson, Daniel 1992: 'Some Considerations on Arithmetical Truth and the ω-rule'. In Detlefsen 1992, pp. 94–138.
——1996: 'Arithmetical Truth and Hidden Higher-order Concepts'. In Hart, 1996, pp. 203–24.
Kaye, Richard 1991: *Models of Peano Arithmetic*. Oxford: Clarendon Press.
Ketland, Jeffrey 1999: 'Deflationism and Tarski's Paradise'. *Mind*, 429, pp. 69–94.
——2005: 'Deflationism and the Gödel Phenomena: Reply to Tennant'. *Mind*, 453, pp. 75–88.
Loewer, Barry 2005: 'On Field's *Truth and the Absence of Fact*: Comment'. *Philosophical Studies*, 1, pp. 59–70.
McGee, Vann 1997: 'How We Learn Mathematical Language'. *Philosophical Review*, 1, pp. 35–68.
McGuinness, Brian, and Gianluigi Oliveri (eds.) 1994: *The Philosophy of Michael Dummett*. Dordrecht: Kluwer.

 © Waxman 2017

Pedersen, Nikolaj Jang Lee Linding, and Marcus Rossberg 2010: 'Open-endedness, Schemas and Ontological Commitment'. *Noûs*, 44(2), pp. 329–39.

Quine, Willard Van Orman 1986: *Philosophy of Logic*. Cambridge, MA: Harvard University Press.

Rayo, Agustín, and Gabriel Uzquiano 1999: 'Toward a Theory of Second-Order Consequence'. *Notre Dame Journal of Formal Logic*, 40(3), pp. 315–25.

Shapiro, Stewart 1991: *Foundations Without Foundationalism: A Case for Second-Order Logic*. Oxford: Oxford University Press.

——1998: 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy*, 10, pp. 493–521.

——2001: 'Systems Between First- and Second-Order Logic'. In Gabbay and Guenthner 2001, pp. 131–88.

——2002: 'Deflation and Conservation'. In Horsten and Halbach 2002, pp. 103–28.

Smorynski, Craig 1977: 'The Incompleteness Theorems'. In Barwise 1977, pp. 821–65.

Tennant, Neil 2002: 'Deflationism and the Gödel Phenomena'. *Mind*, 443, pp. 551–82.

© Waxman 2017

Able to Do the Impossible

Jack Spencer
Massachusetts Institute of Technology
jackspen@mit.edu

According to a widely held principle—the poss-ability principle—an agent, S, is able to ϕ only if it is metaphysically possible for S to ϕ. I argue against the poss-ability principle by developing a novel class of counterexamples. I then argue that the consequences of rejecting the poss-ability principle are interesting and far-reaching.

There is a principle, widely accepted and often taken for granted, that connects abilities and metaphysical possibilities. I call it the *poss-ability principle*:

If S is able to ϕ, then it is metaphysically possible for S to ϕ.

The poss-ability principle states that each of an agent's abilities must be witnessed by a possible exercising thereof; that is, if S is able to ϕ, then there must be some possible world in which S (exercises her ability and) ϕ-s.[1]

As we shall see, quite a lot turns on whether the poss-ability principle is true. The poss-ability principle is a premiss in some of the main arguments for incompatibilism, a consequence of the leading analyses of ability, and often an unstated but critical assumption working in the background. The thesis that an object is visible (or knowable) only if there is some possible world in which the object is seen (or known) relies, implicitly, on the poss-ability principle, as does the widely held thesis that 'ought' implies 'it is metaphysically possible that'.

In this paper I argue against the poss-ability principle by introducing and developing a novel class of counterexamples. I claim that an agent might be able to do what it is metaphysically impossible for her to do. In fact, somewhat more surprisingly, I claim that an agent might be able to do what it is metaphysically impossible to do *tout court*.

[1] Possible worlds are maximally complete possibilities. I assume that incomplete metaphysical possibilities are always contained within complete metaphysical possibilities. So: if there is some possibility in which S ϕ-s, then there is some possible world in which S ϕ-s.

Mind, Vol. 126 . 502 . April 2017 **© Spencer 2016**
doi:10.1093/mind/fzv183 Advance Access publication 16 September 2016

I am not, to be sure, the first philosopher to allege that there are counterexamples to the poss-ability principle. Descartes famously claimed that God's abilities were unconstrained by metaphysical possibility;[2] the early Leibniz, though less radical than Descartes, also thought that God had abilities that could not be exercised;[3] and more recently there has been some discussion of whether the poss-ability principle is true of agents who are situated unusually vis-á-vis time (for example, backwards time-travellers[4] or agents whose four-dimensional bodies form a closed causal loop[5]). I think that considering deities and temporally unusual agents can be instructive, since it helps to mitigate the knee-jerk impulse in favour of the poss-ability principle, but I will for the most part be putting these sorts of extraordinary agents aside. I am interested in a class of counterexamples that arise, not owing to supernatural powers or strange temporal predicaments, but owing to another reason altogether. If the counterexamples that I develop in this paper are successful, then even ordinary human beings might be counterexamples to the poss-ability principle. In fact, I will go on to claim that one of the agents who serves as a counterexample to the poss-ability principle is *you*.

This paper is divided into two sections. First, I argue that we should reject the poss-ability principle, and then I argue that the consequences of rejecting the poss-ability principle are interesting and far-reaching.

1. The case against the poss-ability principle

I begin by identifying a class of cases—*G-cases*—that have an interesting and unexplored structure. I think that G-cases are straightforward

[2] Descartes seems to think that God is able to will the falsity of any truth, even necessary truths, even logical truths; he writes in a 1630 letter to Mersenne that God was 'just as free to make it not true that the radii of a circle are unequal as he was free not to create the world' (AT 1:152, CSMK 25). For further discussion, see Conee (1991), Frankfurt (1977), and Kaufman (2002).

[3] For discussion, see Adams (1994) and Lin (2012).

[4] Fara (2010) and Lewis (1976), among others, argue that backwards time-travellers are counterexamples to the poss-ability principle; they think that although it is metaphysically impossible for an agent to (permanently) kill her younger self, still a backwards time-traveller might be able to do so. For discussion, see Skow (MS), Vihvelin (1996), and Vranas (2010). As I say below in note 13, I think that there is reason to doubt that time-travellers are able to (permanently) kill their younger selves.

[5] For an interesting discussion of agents whose four-dimensional bodies form a closed causal loop, see Kment (MS).

 © Spencer 2016

counterexamples to the poss-ability principle. In a G-case, G is able to ϕ, I claim, even though it is metaphysically impossible for G to ϕ.

My argument against the poss-ability principle does not rely entirely on an appeal to intuition, however. It is also undergirded by a bit of theory. There is clearly *some* interesting and intimate relationship between abilities and metaphysical possibilities, and, after having introduced G-cases, I will put forward what I see as the correct account of this relationship. I claim that the relationship between abilities and metaphysical possibilities is not primarily metaphysical (as it is standardly thought to be), but primarily epistemological. I think that once we appreciate the epistemological relationship between abilities and metaphysical possibilities, (1) we will see what is right about the poss-ability principle, since there is something right about it; (2) we will see what is wrong about the poss-ability principle, since there is something wrong about it; and (3) we will come to expect counterexamples to the poss-ability principle of exactly the G-case variety.

There are many ways to construct G-cases. The simplest take place in deterministic universes and rely on the compatibilistic assumption that agents might be able to do what they are determined not to do. But there are also G-cases that take place in indeterministic universes. Ultimately I think that compatibilists *and* incompatibilists should agree that the poss-ability principle is false.

1.1 The structure of a G-case

More important than any particular G-case is the structure common to them all. I thus begin with the structure and then offer a particular G-case that nicely exemplifies the structure.

In a G-case, one possible world stands apart from the rest of the worlds vis-á-vis the prospects of ϕ-ing. (Strictly speaking, it need not be a single possible world; it could be a class, even an infinite class, of worlds, but the point is easiest to make if we imagine a single world being special.) If ϕ-ing occurs anywhere in modal space, it occurs exclusively at the one possible world. Agents in the one possible world might be able to ϕ, but agents in any other possible world clearly and demonstrably are unable to ϕ. Since there are unexercised abilities, the question arises whether an agent might have the unexercised ability *to* ϕ. And I think that an agent might. Thus the blueprint for a G-case: G has the unexercised ability to ϕ in the one possible world, and lacks the ability to ϕ in every other possible world. There is one possible world in which G is able to ϕ, but there are no possible worlds in which G (exercises her ability and) ϕ-s.

 © Spencer 2016

In §§1.3 and 2.6, we will look at other G-cases. For now, I want to focus on a particularly simple example:

> *Simple G*: Suppose that determinism is true. Let h be the complete specification of the initial conditions of the universe. Let l be the complete specification of the deterministic laws of nature. Let $h \wedge l$ be their conjunction. Suppose that G has not, does not, and will not believe that $h \wedge l$. G never finds herself reading a book or listening to a radio programme about the initial conditions or the laws of nature; G was home from school and sick with the flu on the day that her physics teacher covered the initial conditions and the laws of nature in class, and the physics teacher never bothered to go over the material again. We may suppose that it is fairly common knowledge in G's community that $h \wedge l$, that matriculating high school seniors are expected to know that $h \wedge l$, that many of G's classmates know that $h \wedge l$, and that G is one of the brightest students in her class. The proposition *that* $h \wedge l$ does not exceed G's cognitive wherewithal, either in length or in complexity, and there are no special obstacles preventing G from forming the belief.

Compatibilists, one and all, will agree that G is able to *believe* that $h \wedge l$. I make a stronger claim. I claim that G is able to *truly believe*—or better still—that G is able to *know* or *learn* that $h \wedge l$. G has a *factive* ability.

What are factive abilities?

Factive abilities are a subclass of object-dependent abilities. Object-dependent abilities are abilities to perform object-dependent actions, such as *kissing the Blarney Stone* or *seeing the Statue of Liberty*. Object-dependent actions and abilities are common and familiar. Most of us are able to see the Statue of Liberty, for example, but blind people are not. My sister has an object-dependent ability that I lack: unlike me, she is able to get her dog, Pogo, to roll over.

Factive abilities are object-dependent abilities that depend on facts. Just as the ability to see the Statue of Liberty depends on the existence of the Statue of Liberty, so the ability to know that p depends on it being true that p (or in the language of existence: depends on the existence of the fact that p). Like object-dependent abilities generally, factive abilities are common and familiar. Just think of all of the things that you are able know, learn, discover, and remember!

Factive abilities divide modal space. There are the worlds at which the fact obtains, and there are the worlds at which the fact does not obtain. The worlds at which the fact obtains are the worlds that stand out vis-á-vis the prospects of performing the factive action.

© Spencer 2016

Take $h \wedge l$, for instance. Agents in $h \wedge l$-worlds might know that $h \wedge l$, but agents elsewhere in modal space clearly do not know that $h \wedge l$—knowledge is factive. Similarly, agents in $h \wedge l$-worlds might have the ability to know that $h \wedge l$, but agents elsewhere in modal space clearly lack such an ability—only truths are knowable.

Since there are unexercised abilities, the question arises whether an agent might have the unexercised ability *to know that* $h \wedge l$. And I think that an agent might. In fact, in the vignette above, I think that G has the unexercised ability to know that $h \wedge l$. G is able to know that $h \wedge l$, I think, no less than her fellow classmates. (If G was unable know that $h \wedge l$, then G should not have been enrolled in the physics class to begin with!)

If G is able to know that $h \wedge l$, then we have a counterexample to the poss-ability principle, for there is no possible world in which G knows that $h \wedge l$. A world at which G knows that $h \wedge l$ would have to be a world at which $h \wedge l$ is true. But because the laws of nature are deterministic, every world at which $h \wedge l$ is true is a duplicate of the actual world, and thus every world at which $h \wedge l$ is true is a world at which G fails to come to know that $h \wedge l$. So, while there are worlds at which G is *able* to know that $h \wedge l$, there are no worlds at which G (exercises her ability and) knows that $h \wedge l$.

Of course, *Simple G* is a strange, somewhat artificial example, and such examples should not be asked to pull much philosophical weight. Taken by itself, *Simple G* would be a fairly unconvincing case against the poss-ability principle. But G-cases are only a part of my argument. The other part is a non-standard proposal about the relationship between abilities and metaphysical possibilities, and the argument gathers force only when the two parts act in tandem. I thus want to switch gears and say a bit about why G-cases arise. To that end, we need to broach the epistemology of ability.

1.2 Why do G-cases arise?

Abilities are not easy to detect. It is, in general, much easier to detect the actions that spring from an agent's abilities than the agent's abilities themselves. For that reason, in our epistemology of ability, we use inference to the best explanation, inferring from facts about what agents *do* to conclusions about what agents are *able* to do. The best evidence that an agent is able to ϕ is the agent's repeatedly attempting to ϕ and repeatedly succeeding—Michael Jordan's attempts to slam-dunk a basketball are repeatedly successful, so, by inference to the best explanation, Michael Jordan is able to slam-dunk a basketball. The

 © Spencer 2016

best evidence that an agent is unable to ϕ is the agent's repeatedly attempting to ϕ and repeatedly failing—my attempts to slam-dunk a basketball are repeatedly unsuccessful, so, by inference to the best explanation, I am unable to slam-dunk a basketball.[6]

Keep in mind the order of explanation. It would be wrong to think that the outcomes of our attempts are explanatorily prior to our abilities (as if first someone successfully tries to slam-dunk a basketball and only thereafter and therefore is able to jump so high!). Successful or unsuccessful attempts are not what *make* agents able or unable. Rather, it is the other way around. Our abilities explain the outcomes of our attempts. Michael Jordan's attempts to slam-dunk a basketball are successes, and mine are not, *because* Michael Jordan is able to slam-dunk a basketball, and I am not.

But abilities are hard to detect. You cannot just look at someone and thereby ascertain what she is, and is not, able to do. So the epistemology of ability runs in the opposite direction of the metaphysics. The metaphysics follows the order of explanation: we use what the agent is, and is not, able to do to explain the outcomes of the agent's attempts. The epistemology follows the order of indication: the outcomes of the agent's attempts indicate what the agent is, and is not, able to do. Successful attempts indicate ability; failed attempts indicate inability.

The method of determining what agents are able to do by looking to the outcomes of their attempts would suffer from a poverty of stimulus if we restricted our attention to present or actual attempts, so we take into consideration non-present and non-actual attempts as well. To keep track of things, we index abilities to times and worlds. (After all, what an agent is able to do changes over time and across possibilities.) The question that we want to answer is whether the agent is presently and actually able to ϕ. Our strategy for answering the question is to look at the agent's attempts to ϕ—both present and non-present attempts; both actual and non-actual attempts—and extrapolate from the outcomes of the attempts to our conclusion.

Of course, not all of an agent's attempts are on an evidential par. Some attempts are epistemically relevant, but others are not. To dramatize the point, suppose that in the future I sprout wings.

[6] *Terminological note*: I use the word 'attempt' in a broad sense to mean circumstance in which the would-be ability might reveal itself in action. Attempts, in my sense, need not be effortful or purposeful. Walking outside might be an attempt, in my sense, to sneeze. Reluctantly attending history class might be an attempt, in my sense, to learn that James Madison was the fourth president of the United States.

 © Spencer 2016

Making the most of my new-found appendages, I repeatedly attempt to fly and repeatedly succeed. Such attempts are, indeed, repeated and successful attempts by me to fly. But everyone recognizes that my future (winged) attempts to fly do not provide good evidence about what I (being unwinged) am presently and actually able to do. My (winged) attempts to fly are not, as I will put it, *representative* attempts.

What are the representative attempts? Hold fixed the facts that together determine whether S is presently and actually able to ϕ, and consider the attempts by S to ϕ done at times and worlds at which those facts obtain—*those* are the representative attempts.[7] The attempts by S to ϕ that are representative of whether S is presently and actually able to ϕ are those attempts done by versions of S that are neither enhanced nor diminished (as compared to the present and actual version of S), in circumstances that are neither more nor less favourable to S ϕ-ing (as compared to the present and actual circumstances).

The epistemology of ability goes via the outcomes of the representative attempts. The outcomes of the unrepresentative attempts carry little or no information; we can ignore them.[8] It is the outcomes of the representative attempts that reveal what an agent is, and is not, able to do. Successful representative attempts indicate ability; failed representative attempts indicate inability.

There is a point to make here about the relationship between abilities and counterfactuals. Counterfactuals, as I see them, are epistemological expediters: they streamline the epistemology of ability by providing a quick and, for the most part, reliable way of identifying representative attempts.

We want to know whether an agent is presently and actually able to ϕ. So consider the closest possible worlds in which the agent attempts to ϕ. What are the outcomes of the agent's *closest* attempts? Are they successes or failures? Counterfactuals draw our attention to the outcomes of the agent's closest attempts. And there is, indeed, a robust correlation: for the most part, an agent is able to ϕ just if

[7] Controversies about abilities thus re-emerge as controversies about representative attempts. For example, compatibilists and incompatibilists will disagree about whether representative attempts by S to ϕ must occur in worlds with the same past and laws as the actual world.

[8] Unrepresentative attempts are not always epistemically irrelevant. If S ϕ-s in less favourable circumstances, that may be good evidence that S is presently and actually able to ϕ. Thanks to Eden Lin for discussion on this point.

 © Spencer 2016

her closest attempts are successes. I believe, however, that this correlation is indirect. Abilities and close successes are correlated, not because of some direct link between them, but rather because, for the most part, close attempts are also representative attempts. An agent's closest attempts are typically done by versions of the agent that are neither enhanced nor diminished, in circumstances that are neither more nor less favourable. Thus, close attempts, which are often easier to locate than representative attempts, can serve as useful proxies for representative attempts.

The indirectness of the correlation between abilities and close successes explains why the counterfactual test, although often reliable, occasionally misfires. When the closest attempts are representative attempts, the counterfactual test works fine. But when the closest attempts are unrepresentative, the test makes false predictions. There are two sorts of problem.

Problem 1: *The outcomes of the closest attempts might differ from the outcomes of the representative attempts.* Consider a finkish agent, for instance, me. I have puny biceps, and I am hopelessly unable to do ten chin-ups. But suppose that I have a fairy angel who hates to see me fail. If I tried to do ten chin-ups, my fairy angel would intervene, temporarily beef up my biceps, and my attempt would therefore be a success. Let us stipulate that the closest worlds in which I attempt to do ten chin-ups are worlds in which my fairy angel intervenes and supplements my strength. Good for me, I suppose, for counterfactually doing so many chin-ups! But my (beefed-up) attempts to do ten chin-ups are clearly unrepresentative of what I (being puny) am presently and actually able to do. It may be true that if I tried to do ten chin-ups, I would succeed. It may be true that if I tried to do ten chin-ups, I would magically and temporarily gain the ability to do ten chin-ups. But, despite all of that, I am not presently and actually able to do ten chin-ups. My closest attempts are successes, but my closest attempts are not representative.

There are representative attempts by me to do ten chin-ups, of course; they are attempts done by *unenhanced* versions of me, that is, versions of me that have my same puny biceps, and they are, for the most part, failed attempts. As I said, I am not able to do ten chin-ups.

If there are representative attempts, then we can mitigate, and even eliminate, the difference between the outcomes of the closest attempts and the outcomes of the representative attempts by supplementing the antecedent of the counterfactual. We began with the first counterfactual test: if S attempted to ϕ, then S would (or might) succeed. We do

 © Spencer 2016

better by using the second counterfactual test: if S remained intrinsically very much the same and attempted to ϕ, then S would (or might) succeed. The second counterfactual test is better than the first, because while it is frequently the case that the closest attempts by S to ϕ are representative, it is even more frequently the case that the closest attempts by a not-too-intrinsically-dissimilar version of S are representative. For example, if the closest worlds in which I try to do ten chin-ups are worlds in which my fairy angel supplements my strength, then the first counterfactual test delivers the wrong results, but the second counterfactual test might still get it right.

The second counterfactual test is still imperfect, however, and for the same basic reason: the closest attempts by a not-too-intrinsically-dissimilar version of the agent are not necessarily representative of what the agent is presently and actually able to do. Imagine that my fairy angel operates not by changing my intrinsic nature but by altering my circumstances. When I try do to ten chin-ups, my fairy angel leaves my puny biceps alone, but produces a propitious gust of wind. My (wind-assisted) attempts are no more representative than are my (beefed-up) attempts.

We can continue in this fashion, supplementing the antecedent bit by bit, doing an ever better job of ensuring that the closest, antecedent-satisfying attempts are representative attempts, and thus improving the counterfactual test. For example, we can do better still by using the third counterfactual test: if S remained intrinsically very much the same, and the circumstances remained very much the same, and S attempted to ϕ, then S would (or might) succeed. If we are willing to help ourselves to the notion of representativeness, then we can put an end to this Chisholming and cut straight to the chase.[9] Consider the fourth counterfactual test: if S made a representative attempt to ϕ, then S would (or might) succeed.[10]

The fourth counterfactual test ensures that if there are representative attempts, the closest attempts are among them. This is a dramatic improvement: it eliminates any potential difference between the outcomes of the closest attempts and the outcomes of the representative

[9] Whether it is legitimate to appeal to representativeness depends on our ambitions. If we seek a reductive analysis of ability, it may be illegitimate, since representativeness should probably itself be understood in terms of ability. I seek a reliable test for abilities, so my appeal to representativeness is not illegitimate.

[10] The fourth counterfactual test resembles Lehrer's (1976) semantics for 'can'.

 © Spencer 2016

attempts. In my opinion, the fourth counterfactual test is (qua test of ability) as good as a counterfactual test can be.

But even the fourth counterfactual test is imperfect. The problem is not a potential mismatch between the outcomes of the closest attempts and the outcomes of the representative attempts. Rather, the problem is that there might not be any representative attempts at all.

Problem 2: *There might not be any representative attempts.* There are at least two ways in which it could turn out that there are no representative attempts. First, the agent might be unable to try. Consider a snail. There are no representative attempts by the snail to prove the Riemann hypothesis. To imagine the snail even attempting to prove the Riemann hypothesis requires imagining a significantly enhanced version of the snail. The same might go for me in a coma. It might be impossible for someone in a coma to attempt to run. If so, and if I am indeed in a coma, then there will be no representative attempts by me to run.

Second, it might be that all of the agent's attempts are done in circumstances that are, as compared to the present and actual circumstances, considerably less favourable. This brings me back, finally, to G and her factive abilities.

It may be helpful to start with a temporal analogy. F is considering whether to visit his grandmother. After some deliberation, he decides not to. As of Monday night, let us suppose, F has never attempted to visit his grandmother, and let us suppose, moreover, that F's grandmother dies on Monday night. If, on Tuesday, F attempts to visit his grandmother, then his attempt will be a failure; his grandmother is not there to be visited. But F's attempt on Tuesday to visit his grandmother is not representative of whether he was, on Monday, able to visit his grandmother, for the circumstances on Tuesday are considerably less favourable to F visiting his grandmother than were the circumstances on Monday.

I think that what goes for F in the temporal case goes equally for G in the modal case. To keep things simple, suppose that G has not, and will not, make an attempt to know that $h \wedge l$, where an attempt would be something like asking someone or looking it up online.[11] There are,

[11] Since the simplest G-cases are ones in which G makes no attempt to know that $h \wedge l$, one might wonder whether G is able to attempt to know that $h \wedge l$. We can see that G is able to attempt to know that $h \wedge l$ by looking at slightly more complicated G-cases, in which G makes one or more representative attempts:

> *Mishearing G*: G asks her professor about the laws of nature and initial conditions. Her professor tells her, '$h \wedge l$', but G mishears, and thus comes to believe a nearly true but false proposition.

 © Spencer 2016

of course, possible worlds in which G attempts to know that $h \wedge l$, but all of them are worlds in which $h \wedge l$ is false, and worlds in which $h \wedge l$ is false are considerably less favourable to G knowing that $h \wedge l$ than are worlds in which $h \wedge l$ is true. Thus, modal space contains many attempts by G to know that $h \wedge l$, and all of them are failures; but none of the attempts are representative attempts, and therefore none of the attempts are epistemically helpful. We learn little about whether F was, on Monday, able to visit his grandmother by learning that he fails to visit his grandmother on Tuesday, and we learn little about whether G is presently and actually able to know that $h \wedge l$ by learning that G fails to come to know that $h \wedge l$ in worlds in which $h \wedge l$ is false. In general, failures to perform an object-dependent action in circumstances in which the object is *absent* tell us little or nothing about whether the agent has the object-dependent ability in circumstances in which the object is *present*.

There is, I believe, a true principle, in the vicinity of the poss-ability principle, which connects abilities, attempts, and actions. I call it the *revealing principle*:[12]

> If S is able to ϕ, and there are enough representative attempts by S to ϕ, then at least one of the representative attempts is a success.

(*Question*: Why *enough* representative attempts, rather than *any*? *Answer*: Because there are fluky successes and fluky failures. An agent who is able to ϕ might attempt to ϕ, and fail, not because of inability, but simply because of bad luck. *Question*: How many attempts are *enough*? *Answer*: Enough is a matter both of number and diversity. To ensure that the outcomes are not unduly biased by the circumstances, the attempts must occur in qualitatively diverse circumstances. And to avoid the problem of very bad luck, the attempts must occur in different possible worlds. A single representative attempt can end in fluky failure, so, by a stroke of very bad luck, an agent who is able to ϕ might make any number of representative attempts, and *all* of

---

In *Mishearing G*, G is able to attempt to know that $h \wedge l$, attempts to know that $h \wedge l$, but fails to come to know that $h \wedge l$, as her attempt ends in fluky failure. Thanks to Ned Hall for discussion on this point.

[12] I am actually inclined towards a stronger principle, the long-run principle: if there are enough representative attempts by S to ϕ, then S is able to ϕ iff the long-run proportion of S's representative successes to representative attempts exceeds some minimal threshold. (Different abilities require different thresholds. I am not able to guess the outcome of a coin flip, although I guess right about half of the time; on the other hand, I am able to make a three-point basket, even though I shoot less than 20% from three-point range.) The long-run principle entails the revealing principle.

 © Spencer 2016

her attempts might end in fluky failure. But bad luck cannot persist across worlds. If the agent is able to ϕ, and her representative attempts are scattered across sufficiently many worlds, then it is not possible that *all* of her representative attempts end in fluky failure.)

I accept the revealing principle.[13] If an agent has an ability, then, given enough representative attempts, the ability will reveal itself in action. But to get from the revealing principle to the poss-ability principle, we need an additional premiss. Specifically, we need the premiss that for any S and any ϕ, modal space always contains enough representative attempts by S to ϕ. And it is *this* premiss that I deny. Modal space often contains, but does not *always* contain, enough representative attempts. The revealing principle thus leaves a blind spot, and I claim that within this blind spot lie cases of both ability and inability. Modal space does not contain enough representative attempts by the snail to prove the Riemann hypothesis, and the snail is unable to prove the Riemann hypothesis. On the other hand, modal space does not contain enough representative attempts by G to know that $h \wedge l$, and G is able to know that $h \wedge l$.[14]

On this conception of the relationship between abilities and metaphysical possibilities—the epistemological conception—other possible worlds are like additional rooms in a laboratory: their primary purpose is to supply us with additional representative attempts.[15] If we

[13] Given the revealing principle, we have reason to doubt that time-travellers are able to (permanently) kill their younger selves. For though there may be enough representative attempts by the agent to (permanently) kill her younger self, if auto-infanticide is an impossible crime, none of the attempts are successful.

[14] When modal space fails to contain enough representative attempts, we can sometimes circumvent the problem by going vicarious. Take G. The problem is that none of G's attempts to know that $h \wedge l$ are done in circumstances that are as favourable to G knowing that $h \wedge l$ as are the present and actual circumstances. But there is a second best. We can consider one or more agents who are enough like G and who do indeed make attempts in the favourable circumstances, for example G's classmates. The outcomes of the classmates' attempts are vicariously representative of whether G is presently and actually able to know that $h \wedge l$; they represent what would or might have happened if, per impossible, G had attempted to know that $h \wedge l$ in the favourable circumstances that presently and actually obtain (e.g. had G actually attended the physics lecture that her classmates actually attended). The vicarious attempts are successes—upon attending the lecture, G's classmates come to know that $h \wedge l$—and this suggests that G too is able to know that $h \wedge l$. Going vicarious does not solve the problem entirely, however, as we shall see in §1.3. Even if we supplement the representative attempts with vicarious attempts, still modal space might fail to contain enough representative attempts.

[15] My epistemological conception of the relationship between abilities and metaphysical possibilities is influenced by Kment's epistemological conception of the relationship between causation and metaphysical possibilities; see Kment (2010, 2014).

 © Spencer 2016

want to figure out whether S is able to ϕ, then we could bring her into the lab, put her in a variety of circumstances, and have her repeatedly attempt to ϕ. Extrapolating from the outcomes of her attempts, we could draw an informed conclusion about whether she is, or is not, able to ϕ. Of course, it is not always possible to bring S to the lab for testing, so we face a question: what are we to do when there are not enough actual attempts by S to ϕ to serve as the basis for our extrapolation? Answer: use non-actual attempts. The outcomes of S's non-actual representative attempts give us the same sort of information as do (or would) the outcomes of her actual representative attempts.

The epistemological conception explains both why there are counterexamples to the poss-ability principle and why the counterexamples are rare, strange, and somewhat artificial. Counterexamples to the poss-ability principle are necessarily unexercised abilities. According to the epistemological conception, there are necessarily unexercised abilities for the same reason that there are actually unexercised abilities. The revealing principle is true whether we are considering one world or many. If a set of worlds contains enough representative attempts, then at least one of the representative attempts contained therein is a success. There are actually unexercised abilities because, while S may be able to ϕ, the actual world might fail to contain enough representative attempts. There are necessarily unexercised abilities because, while S may be able to ϕ, the whole of modal space might fail to contain enough representative attempts.

Counterexamples to the poss-ability principle are rare, strange, and somewhat artificial because, in order to construct a case in which modal space fails to contain enough representative attempts, we need to exploit some feature of modal space itself. In the vast majority of cases, modal space contains enough representative attempts—which is why, in the vast of majority of cases, agents' abilities are possibly exercised.

1.3 Constructing other G-cases

Here are five interesting varieties of G-cases.

Close calls. Despite it being metaphysically impossible for G to ϕ, G may get arbitrarily close to ϕ-ing. As G gets closer to ϕ-ing, the intuition that she is able to ϕ grows stronger. Here, for example, is a case that exploits the Gettier condition on knowledge:

> *Gettier G*: Curious about the initial conditions and the laws of nature, G decides to ask her mother, who does indeed know that $h \wedge l$. By a curious twist of fate, however, aliens have randomly chosen one human being to

 © Spencer 2016

abduct for a single night, and G's mother is the unlucky Earthling. For the duration of the abduction, G's mother is replaced by an alien facsimile, who knows neither that h nor that l. G, who justifiably takes the facsimile to be her mother, asks the facsimile, 'What are the initial conditions and the laws of nature?' The facsimile, trying not to blow its cover, takes a wild guess and improbably gets it right, '$h \wedge l$'. On the basis of the facsimile's bad testimony, G comes to the true and justified belief that $h \wedge l$, but G does not know that $h \wedge l$, as G fails the Gettier condition on knowledge. And because G never receives any further evidence, and never changes the basis of her belief, G never comes to know that $h \wedge l$.

G does not, and hence necessarily does not, know that $h \wedge l$, but G seems to be able to know that $h \wedge l$, nevertheless. It is hard to believe that an agent might have the ability to form the true and justified belief that $h \wedge l$, yet lack the ability to know that $h \wedge l$.

Actions that are clearly actions. Perhaps knowing is not an action. There are, however, G-cases that clearly involve actions. Take, for example, the action of *teaching*, where teaching is understood factively:[16]

> *Teacherly G*: F and G are both professors at Tech University. Their offices are adjacent. Students often come by to ask about the initial conditions and the laws of nature. Both F and G know that $h \wedge l$. Students flip a coin to decide which professor to ask—heads, F, tails, G. Perchance, the coins have always landed heads. Thus, while F has taught many students that $h \wedge l$, G has never taught anyone that $h \wedge l$.

It seems to me that G, no less than F, is able to teach students that $h \wedge l$; her ability to teach students that $h \wedge l$ is part of what makes her such a good professor, after all. But G's ability to teach students that $h \wedge l$ is unexercised, and hence necessarily unexercised.

Other facts. To construct a factive G-case we need a proposition, the truth of which necessitates that G does not perform some particular factive action. I have focused on $h \wedge l$, but only for reasons of convenience. Other G-cases involve other facts.[17] One interesting class of G-cases, for example, involves facts concerning actuality.[18] Consider,

[16] Thanks to Kieran Setiya for discussion on this point.

[17] Any proposition that specifies a pocket of determinism, even a pocket contained in an otherwise indeterministic universe, will provide an example. Chalmers (2011) proposes (perhaps unwittingly) an interesting case, which I discuss in §2.5. A different sort of example can be found in Leibniz. If the actual world is the best of all possible worlds, then an agent who does not exercise her ability to know that the world is as good as it could be therefore necessarily does not exercise her ability.

[18] Kearns (2011) also discusses facts concerning actuality, arguing that agents might be responsible for impossibilities.

 © Spencer 2016

to take the most extreme example, the proposition *that the actual world is actual*. This is a maximally contingent proposition, a truth that would be false if the world were different in any way at all. I know that the actual world is actual, as do you. But not everyone knows that the actual world is actual; many have never even considered the matter. The question arises, then, whether an agent might have the unexercised ability to know that the actual world is actual. And I think so:

> *Actual G*: G is an actual person, a competent college student, who, as a matter of fact, will never come to believe that the actual world is actual.

I think that G, in *Actual G*, like most college students, has the unexercised (and therefore necessarily unexercised) ability to know that the actual world is actual.[19]

Let me pause here to address incompatibilists, in particular. Incompatibilists should not be persuaded by *Simple G*, *Gettier G*, or *Teacherly G*, as all of these cases rely on the assumption that an agent might be able to do what she is determined not to do. But I think incompatibilists should find other G-cases persuasive, among them *Actual G*. We can imagine that the actual world is indeterministic, that agents are free from manipulation and coercion, that agents have agent-causal powers and so on. G, in *Actual G*, does not, and hence necessarily does not, come to know that the actual world is actual, but it would be wrong, I think, to say that G is somehow determined not

[19] Some have objected to *Actual G*. They doubt whether we would have been able to think about the actual world had the world been otherwise, and thus suggest that the ability to entertain the proposition that the actual world is actual cannot be unexercised. I disagree. I think that agents might have the unexercised ability to entertain the proposition that the actual world is actual. But a more concessive response can be given. Consider a hybrid of *Actual G* and *Gettier G*, on which G truly believes, but fails to know, that the actual world is actual. In such a case, G is both able to entertain, and able to know, that the actual world is actual, even though it is metaphysically impossible for G to know that the actual world is actual.

A helpful referee raises a different objection. The proposition that the actual world is actual appears to involve a singular thought about the actual world. What is needed for the counterexample to work is a maximally contingent proposition, not the tautology that the world that is actualized is actualized, so 'the actual world' must designate rigidly. Of course, we can know the actual world by description, as the world that is actually actualized. But can we entertain singular thoughts about the actual world? The question of whether we can entertain singular thoughts about the actual world is vexed and interesting, but I think that we can finesse the matter. Let 'the actual world' be a rigidified description, which rigidly designates the world that is actually actualized. The resultant proposition is then maximally contingent, even if it is not singular. The same dialectic then ensues. If G could have the unexercised ability to know this maximally contingent proposition, then we have a counterexample to the poss-ability principle, even by the lights of incompatibilism.

 © Spencer 2016

to come to know.[20] It is *up to* G, I think, in the relevant sense, whether she comes to know that the actual world is actual, even though it is metaphysically necessary that she will not.

Non-factive cases. Whether there are non-factive G-cases depends on what objects there are. All G-cases involve object-dependent abilities, where the existence of the object necessitates that G does not perform the object-dependent action. If objects are sparse and easy to recombine modally, then there might not be any non-factive G-cases. But some philosophers (myself included) believe that objects are plentiful. For instance, consider a large rock in Siberia. Someone inclined by what Karen Bennett (2004) calls 'bazillion thing-ism'[21] thinks that there is a modally fragile object, which is just like the rock, the same size as the rock, the same colour as the rock, co-located with the rock, but different from the rock, because, unlike the rock, it has all of its properties essentially. Suppose that nobody ever sees the large rock, and hence that nobody ever sees this modally fragile object. Is this modally fragile object visible? Might an agent have the ability to see it?

I think that if this modally fragile object exists, then clearly it is a visible object. It is as big as a rock, as brown as a rock, as conspicuous as a rock, and reflects photons just like a rock—indeed, necessarily so! Moreover, agents might have the ability to see it. The fact that this modally fragile object is unseen (and hence necessarily unseen) is not due to a collective lack of ability; it is due, rather, to the fact that modal space affords us so few opportunities to see it.

It is controversial whether these sorts of modally fragile object exist, and I will not assume that they do. Rather, I will continue to focus on factive abilities.

Lonely cases. In a Lonely G-case, G is the only being in all of modal space who is rightly positioned to perform a certain factive action:

> *Lonely G*: Suppose that $h \wedge l$ is true, and hence that determinism is true. G, the only intelligent being in the universe, has lived for millions of years, and in that time developed all of modern science and mathematics. As it happens, G never comes to believe that $h \wedge l$, but this is an historical accident. (Perhaps G never considers the question, or perhaps G makes an uncharacteristic blunder in her calculations and thus comes to a false belief.) We can supply G with all the requisite technology (perhaps G need

[20] Being determined not to ϕ requires more than a proposition whose truth necessitates that the agent does not ϕ, else unexercised abilities would be impossible!

[21] Proponents of bazillion thing-ism include Eklund (2008), Fine (1999), Hawthorne (2006), Johnston (2006), Leslie (2011), and Yablo (1987).

 © Spencer 2016

only type the question into her supercomputer) and make G as able-minded as we like.

The following principle is even weaker than the poss-ability principle:

if S is able to ϕ, then it is metaphysically possible to ϕ.

Lonely G is, however, a counterexample to even this very weak principle. G, in *Lonely G*, is able to do something that not a single being in all of modal space manages to do.

1.4 Summary

Before we reject the poss-ability principle, we need not just a class of alleged counterexamples but also a satisfying explanation for why the poss-ability principle has held such appeal. I claim that by coupling G-cases with the revealing principle, we have what we need. The poss-ability principle has seemed true because counterexamples are rare. Agents' abilities are almost always possibly exercised. Counterexamples appear only when we turn our attention to the cases for which modal space fails to contain enough representative attempts, namely, G-cases.

2. Consequences of rejecting the poss-ability principle

The revealing principle is only slightly weaker than the poss-ability principle. Nevertheless, the consequences of rejecting the poss-ability principle and retreating to the revealing principle are interesting and far-reaching.

2.1 Abilities

The first consequence belongs to metaphysics. An analysis of ability is a specification of what it is for an agent to have an ability. The leading analyses of ability take ability to be analysed in terms of metaphysical possibility; they are, in that sense, *modal* analyses. If we reject the poss-ability principle, then modal analyses of ability fail.

There are two main sorts of modal analysis of ability: counterfactual analyses and restricted possibility analyses. Let me start with counterfactual analyses. According to a counterfactual analysis: S is able to ϕ iff S would (or might) ϕ if S ψ-ed. The most famous counterfactual analysis is the so-called conditional analysis, due to G. E. Moore (1912), according to which S is able to ϕ iff S would ϕ if S tried to ϕ. On the standard Stalnaker-Lewis semantics for counterfactuals, 'S would ϕ if S tried to ϕ' is true iff some or all of the closest worlds in

 © Spencer 2016

which S tries to ϕ are worlds in which S ϕ-s. Similarly, 'S would (or might) ϕ if S ψ-ed' is true iff some or all of the closest worlds in which S ψ-s are worlds in which S ϕ-s.

G-cases make trouble for counterfactual analyses in the obvious way. In a G-case, G is able to ϕ, even though there is no possible world in which G ϕ-s. In a Lonely G-case, G is able to ϕ, even though there is no possible world in which anyone ϕ-s. G-cases thus suggest that there is no counterfactual of the form 'S would (or might) ϕ if S ψ-ed' that is a necessary condition on S's being able to ϕ. An agent might have an ability, even though there is no circumstance in which she would (or might) exercise the ability.

G-cases also make trouble for restricted possibility analyses. There is a sense of 'can'—often called the 'can' of ability—for which the following biconditional is uncontroversially true: S is able to ϕ iff S can (in the special sense) ϕ. The 'can' of ability denotes a particular grade of possibility, which I shall here label *agentive possibility*. We thus have three equivalent ways of speaking: S is able to ϕ iff S can (in the special sense) ϕ iff it is agentively possible for S to ϕ.

According to the restricted possibility analysis, agentive possibility is a restricted grade of metaphysical possibility, where *x*-possibility is a restricted grade of *y*-possibility iff to be *x*-possible is to be *y*-compossible with a certain set of facts. The paradigm is nomic possibility, which is a restricted grade of metaphysical possibility: to be nomically possible is to be metaphysically compossible with the nomic facts. If agentive possibility is likewise a restricted grade of metaphysical possibility, then there is some (perhaps time-, individual- and context-relative) set of facts—we may call them the *agentive facts*—such that to be agentively possible is to be metaphysically compossible with the agentive facts. For S to be able to ϕ, according to the restricted possibility analysis, is for it to be metaphysically compossible with the agentive facts that S ϕ-s.

If any conception of ability deserves to be called the prevailing view, it is the restricted possibility analysis.[22] But if the poss-ability principle

[22] See, for example, Lewis (1976) and van Inwagen (1983). Mele says, 'Philosophers happy to talk in terms of possible worlds will say that an agent at a world W is [able] to A at t if and only if she (or a counterpart) A-s at t in some relevant possible worlds … One way to see the disagreement about incompatibilists and compatibilists about determinism and being able to do otherwise is as a disagreement about what worlds are relevant. According to incompatibilists, all and only worlds with the same past and natural laws as W are relevant; they hold the past and the laws fixed. Compatibilists disagree' (Mele 2003, p. 451). Vetter treats the restricted possibility of analysis as given, and 'propose[s] that we extend the restricted possibility of abilities to dispositions' (Vetter 2014, p. 133).

 © Spencer 2016

is false, then so too is the restricted possibility analysis. After all, no matter what the agentive facts are, it is not metaphysically compossible with *them* that G knows that $h \wedge l$![23]

2.2 The ability to do otherwise

The second consequence also belongs to metaphysics. One of the most important arguments for incompatibilism is the Consequence Argument.[24] If the poss-ability principle is false, however, then the Consequence Argument is unsound.

The Consequence Argument, as it is standardly presented, involves a sentential operator, 'N', and two rules of inference that govern its application, Rule (α) and Rule (β). According to Rule (α): $\Box p \vdash \mathrm{N}p$. (In English: if it is metaphysically necessary that *p*, then nobody is able to render it false that *p*.[25]) According to Rule (β): $\mathrm{N}p$, $\mathrm{N}(p \rightarrow q) \vdash \mathrm{N}q$.[26] (In English: if nobody is able to render it false that *p*, and nobody is able to render it false that *p* materially implies that *q*, then nobody is able to render it false that *q*.) Whereas Rule (β) is famously controversial, many regard Rule (α) as a virtual

[23] Objection: Even if existing modal analyses of abilities entail the poss-ability principle should thus be rejected, couldn't we formulate a new modal analysis that did not entail the poss-ability principle? Reply: Yes. One option is a counterpart-counterpart theory, according to which S is able to ϕ iff in the right circumstances a counterpart of S would (or might) perform an action that is a counterpart of ϕ-ing. Given the right sort of counterpart relation—if, for example, believing that $h \wedge l$ is a counterpart of knowing that $h \wedge l$—the counterpart-counterpart theory may evade the problem of G-cases. But we should still accept that G-cases are counterexamples to the poss-ability principle, and we should still accept the epistemological conception. If by accepting the counterpart-counterpart theory we evade the problem of G-cases, we do so only by increasing the pool of representative attempts. The counterpart-counterpart theory is extensionally adequate (if it is) only because (i) the revealing principle is true, and (ii) since attempts by counterparts of S to do actions that are counterparts of ϕ-ing now count as representative of whether S is presently and actually able to ϕ, modal space always contains enough representative attempts. Thanks to Sarah-Jane Leslie and Steve Yablo for discussion on this point.

[24] See, for example, Ginet (1966) and van Inwagen (1983).

[25] I gloss N in terms of ability, since I am concerned with interpretations of N on which N(S does not ϕ) entails that S is not able to ϕ. As a helpful referee points out, however, one could adopt alternative understandings of N. For instance, in the quoted passage below, van Inwagen glosses N in terms of choice: N*p* iff nobody has a choice about whether *p*. I think that van Inwagen, himself, even when he glosses N in terms of choice, thinks that N(S does not ϕ) entails that S is not able to ϕ, since he thinks that freedom is the ability to do otherwise. But one could sever the connection between abilities and choices. The result would be an interesting conception of choice, on which, though some agent is able to render it false that *p*, nobody has a choice about whether *p*.

[26] Here and throughout, '→' and '↔' are the material conditional and biconditional, respectively.

 © Spencer 2016

truism. Kadri Vihvelin (2011, §5) says, 'Rule Alpha is uncontroversial'; Alicia Finch and Ted A. Warfield (1998, p. 517) say, 'Rule Alpha is surely unobjectionable'. Peter van Inwagen himself says,

> I do not see how anyone could reject Rule (α). If (α) is invalid, then it could be true that someone has a choice about what is necessarily true. Hardly anyone besides Descartes has been willing to concede such a capacity to God. No one, so far as I know, has ever suggested that human beings could have a choice about what is necessarily true. (van Inwagen 1983, p. 96)

I, however, do reject Rule (α). I also reject van Inwagen's intimation that it would take a creature of godly powers to invalidate Rule (α). I think that Rule (α) is invalid, even when restricted to human beings. G-cases are counterexamples. It is necessarily false that G knows that $h \wedge l$, but G is able to know that $h \wedge l$. Hence, for any interpretation of 'N' on which N(S does not ϕ) implies that S is unable to ϕ, Np does not follow from $\Box p$. Moreover, if N($p \rightarrow q$) follows from $\Box q$, as presumably it does, then G-cases are also counterexamples to Rule (β).[27]

The failure of the poss-ability principle implies, then, that the Consequence Argument, at least as it is presently formulated, is unsound.[28] As menacing as this conclusion may appear, I suggest that incompatibilists see it less as a menace than as a challenge. In my view, everyone should reject the poss-ability principle, compatibilists and incompatibilists alike. And I suspect that the Consequence Argument can be formulated in a way that does not rely on the poss-ability principle. But the failure of the poss-ability principle does present a challenge: it forces incompatibilists to say more about how, exactly, determinism poses a threat to freedom. (I here understand freedom as the ability to do otherwise.)

According to a long and storied tradition, the threat of determinism is none other than the threat of necessity. On this way of seeing things, freedom requires that it be contingent that we act as we do;[29] determinism implies that it is necessary that we act as we do; and the

[27] There are other ways to formulate the Consequence Argument. In Finch and Warfield's (1998) presentation, for instance, the Consequence Argument relies on a single rule of inference, namely, Rule (β-2): Np, $\Box(p \rightarrow q) \vdash$ Nq. (Rule (β-2) also features in Lamb's (1977) formulation of the Consequence Argument.) G-cases are counterexamples to Rule (β-2).

[28] I focus on the Consequence Argument, but the poss-ability principle is a premiss in many arguments that leverage metaphysical necessity against freedom. See, for example, Holliday's (2012) Action-Type Argument for incompatibilism, and Pike's (1965) argument for the incompatibility of freedom and foreknowledge.

[29] Frankfurt (1969) argues that freedom does not require that it be contingent that we act as we do. Fara (2008) and Vihvelin (2013) argue against Frankfurt. I am convinced by Fara and Vihvelin.

 © Spencer 2016

fundamental question in the compatibility debate is whether the contingency required by freedom is compatible with the necessity implied by determinism. The Consequence Argument is an attempt to lay bare the incompatibilists' view that the necessity implied by determinism excludes the contingency required by freedom. According to the Consequence Argument: (1) if determinism is true, then, holding fixed the past and the laws, it is metaphysically impossible for agents to act otherwise; (2) nobody has any choice about the past and the laws, so the facts about the past and the laws should be included among the agentive facts; and (3) the contingency that freedom requires is metaphysical contingency holding fixed the agentive facts. From (1), (2) and (3), incompatibilism follows.

There is, of course, a great deal of controversy about (2).[30] But what the failure of the poss-ability principle shows us is that, even putting such controversies aside, (3) is false. The Consequence Argument, at least as it is presently formulated, relies on a mistaken conception of the contingency that freedom requires. The contingency that freedom requires is not metaphysical contingency; hence it is not metaphysical contingency holding fixed some set of facts. Thus the challenge for incompatibilists: why, exactly, does determinism pose a threat to freedom?[31] The threat of determinism cannot simply be the threat of metaphysical necessity (holding fixed some facts), for S might be able to ϕ, even though it is metaphysically necessary that S does not ϕ.

For compatibilists, the failure of the poss-ability principle is interesting, because it brings into focus the question of whether metaphysical contingency matters at all. Does the freedom enjoyed by agents in deterministic universes crucially depend on the past or the laws being metaphysically contingent? Many compatibilists have thought that agents in deterministic universes act freely only if (and only because) the past or the laws are metaphysically contingent.[32] But if we reject the poss-ability principle, then there is room for the opposing view, that agents might act freely, even in the absence of metaphysical contingency—which, in so far as I am inclined toward compatibilism, is

[30] See, for example, Holliday (2012), Huemer (2000), and Lewis (1981)

[31] This question, I think, can be answered, though I suspect that different incompatibilists will answer it differently. A fan of agent causation (see, for example, O'Connor 2000) might think that the ability to do otherwise is the ability to agent-cause yourself to do otherwise, and that the causal necessity associated with determinism excludes the possibility of agent-causation. So, while the ability to ϕ might be compatible with it being metaphysically impossible to ϕ, the ability to ϕ is not compatible with being determined not to ϕ.

[32] See, for example, Lewis (1981), Fischer (2006a), and Vihvelin (2013).

 © Spencer 2016

the sort of view that I would want to defend. Such a view might be labelled an 'actual sequence' conception of the ability to do otherwise.[33] It says that determinism might be true, the past and the laws might be metaphysically necessary, and still agents might act freely.

2.3 'Can'

The third consequence belongs to semantics. In semantics, the 'can' of ability is standardly treated as a possibility modal.[34] Oversimplifying, if the 'can' of ability is a possibility modal, then there is a space of worlds, and a way of selecting a subset thereof, such that $[\![\text{can } \phi]\!]$ is true iff there is at least one world among the selected subset at which $[\![\phi]\!]$ is true. If the poss-ability principle is false, then one of two consequences follows. Either we must abandon the project of treating the 'can' of ability as a possibility modal[35] or the space of worlds that we use for semantics must include worlds that are not metaphysically possible.[36]

2.4 'Ought' and moral responsibility

The fourth consequence belongs to metaethics. There are several alleged connections between moral notions and metaphysical modality. I believe, however, that these connections are indirect; they are mediated by a direct connection between moral notions and ability. And since the poss-ability principle is false, these alleged connections admit of counterexamples.

Consider the following G-case, in which G's failure to ϕ is wilful, deliberate, and morally weighty:

Wilful G: Suppose that $h \wedge l$ is true, and hence that determinism is true. G receives a one-question true-or-false quiz. The question: 'True or

[33] Wolf (1990) seems to have this sort of 'actual sequence' view of the ability to do otherwise. For a defence of an 'actual sequence' conception of moral responsibility, see Fischer (2006b).

[34] See, for example, Krazter (1977, 1981).

[35] See, for example, Vetter's (2013) semantics for 'can', which does not invoke possible worlds.

[36] Note that the usual problems associated with moving from a possible worlds semantics to a possible-and-impossible worlds semantics (see, for example, Berto 2013 and Stalnaker 1996) are exacerbated in the present context, because we are interested in factive abilities. For example, if, at w, the impossible world at which G knows that $h \wedge l$, it is both true and false that G believes that $h \wedge l$, then it would seem to follow that G is able to know a proposition that she does not believe, since there is a relevant (albeit impossible) world in which she does. But G is not able to know propositions that she does not believe.

 © Spencer 2016

false: $h \wedge l$?' As it happens, G's father is up for the PTA Father of the Year Award, which he both yearns for and deserves to win. G knows that her father will lose the award if she fails the quiz. Out of nothing but wilful and bitter teenage rebellion, G intentionally fails the true-or-false quiz. G, who knows full well that $h \wedge l$, incorrectly answers that $h \wedge l$ is false.

Correctly answering that $h \wedge l$ is true is a factive action; it is an action that most of G's classmates perform, an action that, in my view, G is able to perform, yet an action that it is metaphysically impossible for G to perform. Here, where ability and metaphysical possibility part ways, moral notions seem to me to travel together with ability.

I claim that G ought (in the moral sense) to correctly answer that $h \wedge l$ is true, and therefore that G does as she ought not to do.[37] There may be a true version of 'ought' implies 'can'. If so, it is the principle that 'ought' implies 'it is agentively possible that': ought implies *ability*. The principle that 'ought' implies 'it is metaphysically possible that' is false; *Wilful G* is a counterexample. It is metaphysically impossible for G to correctly answer that $h \wedge l$ is true, but G ought (in the moral sense) to correctly answer that $h \wedge l$ is true.

Similarly, I claim that G is morally responsible and blameworthy for failing to correctly answer that $h \wedge l$ is true (or equivalently: G is morally responsible and blameworthy for incorrectly answering that $h \wedge l$ is false). Typically, if an agent ought to ϕ, but cannot, the agent is thereby excused for failing to ϕ. But the fact that it is metaphysically impossible for G to correctly answer that $h \wedge l$ is true does not seem to excuse G for failing to do so. There might be a true version of the principle of possible action.[38] If so, it is the principle of agentively possible action: an agent is morally responsible for failing to ϕ only if it is agentively possible that the agent ϕ. The principle of metaphysically possible action is false; *Wilful G* is a counterexample. It is metaphysically impossible for G to correctly answer that $h \wedge l$ is true, but G

[37] *Objection*: It is not true that G ought to correctly answer that $h \wedge l$ is true. What is true is that G ought not to have the bitter thoughts and desires that lead her to incorrectly answer that $h \wedge l$ is false, but it is metaphysically possible for G to lack those bitter thoughts and desires. *Reply*: Yes, G ought not to have those bitter thoughts and desires, but, in addition, G ought not to do as she does. Consider a psychological duplicate of G, G*, who similarly knows that $h \wedge l$, who similarly knows about her father's desire for the PTA Father of the Year Award, who has the same bitter thoughts and desires as does G, but who, for whatever reason, correctly answers that $h \wedge l$ is true, perhaps unintentionally. It stands to reason that G has done more wrong than G* has. G has more to regret, more to atone for, than does G*. Thanks to Caspar Hare and Steve Yablo for discussion on this point.

[38] The principle of possible action was introduced by van Inwagen (1978).

 © Spencer 2016

is morally responsible and blameworthy for failing to correctly answer that $h \wedge l$ is true.[39]

2.5 The blankables and, in particular, knowability

The fifth and final consequence belongs to philosophical logic. The *blankables* are a subclass of words in English that end '-able' or '-ible'; they include the words 'visible', 'perceptible', 'discoverable', 'thinkable', 'provable', 'fallible', and 'conquerable'.[40] The blankables are to be analysed in terms of ability. A visible object is an object that can, in the ability sense, be seen. A knowable object is an object that can, in the ability sense, be known.

Like ability, the blankables are often analysed in terms of metaphysical possibility. The claim that x is visible is regimented as the claim that it is possible that x be seen: '$\Diamond$Sx' . The claim that p is knowable is regimented as the claim that it is possible that p be known: '$\Diamond$Kp'.

If the poss-ability principle were true, then the modal regimentation of the blankables might be legitimate. But the poss-ability principle is false, and the modal regimentation of the blankables is not just illegitimate, but also a source of philosophical confusion.

Start with *exercisability*. Some have objected to my negative thesis—that G-cases are counterexamples to the poss-ability principle—on the grounds that abilities must be exercisable. But this objection misses its mark. I fully agree that abilities must be exercisable. The crucial question is whether we should understand exercisability in terms of ability or in terms of metaphysical possibility. I say that something might be exercisable even though it is necessarily unexercised. G's ability to know that $h \wedge l$ is a good example.

In general, the distinction between being necessarily unblanked and being unblankable strikes me as neglected but important. I suspect

[39] Whether there are obligations that are impossible to satisfy bears on a number of questions in deontic logic. For example, some have attempted to understand obligation as a restricted grade of metaphysical necessity. Where d is the claim that all obligations are met, obligation is then understood as compossibility with d. It is unclear whether this 'reduction of deontic logic to alethic modal logic', as Anderson (1958) calls it, can succeed, however, if there are impossible obligations. For more on the Andersonian-Kangerian reduction, see Kanger (1957), Lokhorst (2006), and Mares (1992).

[40] There is another class of words that end '-able' or '-ible', which include the words, 'admirable', 'desirable', 'valuable', and 'enviable'. These words, at least on their most natural readings, are of a kind with 'trustworthy' and 'newsworthy'; they are to be analysed, not in terms of ability, but in terms of fittingness or worthiness. To be admirable, for instance, is to be a fitting object of admiration.

 © Spencer 2016

that some perceptible things are necessarily unperceived (I gave a possible example in §1.3); I suspect that some tangible things are necessarily untouched; and I suspect that some knowable things are necessarily unknown. In the remainder of the paper, I want to focus on knowability. The fact that something can be knowable yet necessarily unknown can help to shed light, I think, on two paradoxes of knowability.

Fitch's paradox of knowability. Anti-realists of a Dummettian bent hold that there is a knowability constraint on truth: all truths are knowable.[41] There is a famous argument against the knowability constraint, first published by Frederic Fitch (1963). The argument begins by regimenting knowability as '$\Diamond$K' , and proceeds as follows:

(1) $p \rightarrow \Diamond Kp$.	premiss, the knowability constraint
(2) $q \wedge \neg Kq$	premiss, some truth is unknown
(3) $(q \wedge \neg Kq) \rightarrow \Diamond K(q \wedge \neg Kq)$	from (2), by the knowability constraint
(4) $\Diamond K(q \wedge \neg Kq)$	from (2) and (3), by sentential logic
(5) $\Diamond(Kq \wedge K\neg Kq)$	from (4), by $\Box(K(p \wedge q) \rightarrow (Kp \wedge Kq))$
(6) $\Diamond(Kq \wedge \neg Kq)$	from (5), by the factivity of knowledge

But (6) is false; contradictions are not possibly true. Yet the other premisses are true, and the inferences are valid, so the knowability constraint stands refuted—or so says Fitch's argument.

I think that we can expose a flaw. We should respond to Fitch in the form of a dilemma. If '$\Diamond Kp$' means that it is possible that p be known, then the argument may be sound, and its conclusion may be true, but the argument says nothing about the knowability constraint. The argument is then a refutation of the 'possibly known' constraint on truth: all truths are possibly known. One could propose a 'possibly known' constraint on truth, I suppose, but that was not the anti-realists' proposal. The anti-realists proposed a knowability constraint on truth, and the knowability constraint does not imply the 'possibly known' constraint. A proposition might be knowable yet necessarily unknown. I am suggesting that $K(q \wedge \neg Kq)$ may be one such example.[42]

[41] Dummett wavers on what, exactly, the knowability constraint is supposed to come to; see Dummett (2001, 2009). For a good recent discussion, see Hand (2014).

[42] Fara (2010) also suggests that $K(q \wedge \neg Kq)$ might be knowable but necessarily unknown. Fara does not, however, discuss G-cases, the epistemological conception, or the bearing of the failure of the poss-ability principle on the compatibility debate or on metaethics.

 © Spencer 2016

On the other horn of the dilemma, if '$\Diamond Kp$' is stipulated to mean that p is *knowable*, then the move from (4) to (5) is invalid. The fact that a conjunction is knowable does not imply that there is a possibility in which each of the conjuncts is known.

I myself do not accept the knowability constraint on truth, but my advice to those who do is get clearer about the relationship (and distinction) between being knowable and being possibly known. Fitch's argument does not obviously give us reason to abandon the knowability constraint on truth.

Let me briefly switch from offence to defence. Some who have seen or heard this paper have thought that the argument that I run from the failure of the poss-ability principle to the potential knowability of Fitch propositions should be run *modus tollens* against me. They treat the unknowability of Fitch propositions (i.e. propositions of the form $p \wedge \neg Kp$) as a fixed point, and believe that any theory that implies the knowability of Fitch propositions thus stands refuted. Does rejecting the poss-ability principle necessarily entail that Fitch propositions are knowable?

One argument suggests that it does. Suppose that G is able to know that $h \wedge l$. The truth of $h \wedge l$ necessitates that G does not come to believe that $h \wedge l$, so G (or G+, a suitably enhanced version of G) ought to be able to know that $\Box((h \wedge l) \rightarrow \neg K(h \wedge l))$. By a plausible principle of closure, then, G (or G+) ought to be able to know that $h \wedge l \wedge \neg K(h \wedge l)$. But this is a Fitch proposition, hence the allegedly threatening conclusion: G-cases are counterexamples to the poss-ability principle *only if* Fitch propositions are knowable.

I respond to this objection by taking issue with it at three points. First, the argument that Fitch propositions are always unknowable remains to be seen. The standard argument relies on the poss-ability principle, but I reject the poss-ability principle. At least in the present context, another argument is needed. Second, necessitation is not always epistemically transparent. Although $\neg K(h \wedge l)$ may be necessitated by $h \wedge l$, it need not be scrutable on the basis of $h \wedge l$, so there is no reason to think that G (or G+) ought to be able to know that $\Box((h \wedge l) \rightarrow \neg K(h \wedge l))$. Thus, even if the threatening argument is sound—which I doubt very much—there would still be *some* G-cases that did not lead to the knowability of Fitch propositions, and so the main claims of the paper would stand. But third, and most importantly, if we start from the premiss that Fitch propositions are always unknowable, then we can show that a number of plausible-sounding closure principles fail. Knowability is not closed under

 © Spencer 2016

conjunction introduction; nor it is closed under known implication.[43] And although it is hard to find examples that are completely uncontroversial, if Fitch propositions are always unknowable, then I am inclined to think that knowability also fails to be closed under known necessary implication.[44] I am thus unconvinced by this Fitch-y objection. Perhaps G is able to know a Fitch proposition; and if not, if Fitch propositions are always unknowable, for instance, then G might still be able to know both that $h \wedge l$ and that $\Box((h \wedge l) \rightarrow \neg K(h \wedge l))$. G-cases thus would still be counterexamples to the poss-ability principle.[45]

[43] Here is a counterexample to the closure of knowability under conjunction introduction, assuming that Fitch propositions are always unknowable. Let p be a mathematical truth. Suppose that S is the best mathematician in the history of the world, and knows as much. With an hour of concerted effort, S could prove that p. I assume that, if S is able to prove that p, S is able to know that p. So S is able to know that p. S knows, however, that she will never come to know whether p. She knows that it would take her an hour of concerted effort to determine whether p, and she knows that she will never devote an hour of concerted effort to determining whether p. S also knows that nobody else knows whether p, as she knows that she alone possesses the mathematical prowess needed to determine whether p. So S knows that p is unknown. Knowledge implies knowability, so S is able to know that $\neg Kp$. But if Fitch propositions are always unknowable, then S is not able to know that $p \wedge \neg Kp$, so knowability is not closed under conjunction introduction.

A similar argument shows that knowability is not closed under known implication. Imagine a list containing one thousand mathematical statements: $p_1, \ldots, p_{1000}$. S knows that exactly one statement on the list is true. Suppose that p_i is the lone truth. S is able to prove that p_i, and hence S is able to know that p_i. Moreover, for each p_j, $p_i \neq p_j$, S is able to prove that p_j is false. For every statement on the list, p_n, S knows that she will never come to know to whether p_n; for she knows that it would take her an hour of concerted effort to determine whether p_n, and she knows that she will never devote such time to determining whether p_n. S also knows that nobody else knows whether p_n, as nobody else possesses the requisite mathematical prowess. Performing a simple deduction, S comes to know, for each p_n, that $p_n \rightarrow (p_n \wedge \neg Kp_n)$. In particular, S knows that $p_i \rightarrow (p_i \wedge \neg Kp_i)$. But if Fitch propositions are always unknowable, then S is not able to know that $p_i \wedge \neg Kp_i$, so knowability is not closed under known implication.

[44] I am tempted by a hypothesis, suggested to me by Nilanjan Das, that the only non-trivial operations under which knowability is closed are single-premiss operations. Consider a logically valid deduction from some premisses, $p_1, p_2, \ldots, p_n$ to q. If each premiss, p_i, can be known without q being known, then, since knowability is not closed under conjunction introduction, we can construct a model in which each of the premisses is knowable but the conclusion is not, since there is no accessible world in which the premisses are known together. This give us model-theoretic reason to think that knowability should fail to be closed even under known necessary implication.

[45] Thanks to Nilanjan Das, Katie Elliott, Wes Holliday, Matthew McGrath, Bernhard Salow, Kieran Setiya, James Shaw, Jonathan Vogel, and Ian Wells for discussions of knowability and closure.

 © Spencer 2016

Chalmers's paradox of knowability. A second paradox of knowability has recently been put forward by David Chalmers:

> It is widely believed that for all p, or at least all entertainable p, it is knowable a priori that (p iff actually p). It is even more widely believed that for all such p, it is knowable that (p iff actually p). There is a simple argument against these claims from four antecedently plausible premisses … [Let] 'A', 'E', 'K', '$\Box$', '$\Diamond$' stand for 'Actually', 'Someone entertains', 'Someone knows', 'Necessarily', and 'Possibly', while '$\rightarrow$' and '$\leftrightarrow$' are the material conditional and biconditional. In addition, q is any (entertainable and expressible) proposition that no-one actually entertains, while r is $\neg \mathrm{E}q$, the proposition that no-one entertains q.
>
> (1) $\mathrm{A}r$
>
> (2) $\mathrm{A}r \rightarrow \Box \mathrm{A}r$
>
> (3) $\Box(\mathrm{K}(r \leftrightarrow \mathrm{A}r) \rightarrow (r \leftrightarrow \mathrm{A}r))$
>
> (4) $\Box(r \rightarrow \neg \mathrm{K}(r \leftrightarrow \mathrm{A}r))$
>
> (5) $\neg \Diamond \mathrm{K}(r \leftrightarrow \mathrm{A}r)$
>
> … The conclusion follows from the premisses by classical logic and weak modal logic **K** … From (3) and (4), one can derive $\Box(\mathrm{K}(r \leftrightarrow \mathrm{A}r) \rightarrow ((r \leftrightarrow \mathrm{A}r) \wedge \neg r))$. From (1) and (2), one can derive $\Box(\mathrm{K}(r \leftrightarrow \mathrm{A}r) \rightarrow \mathrm{A}r)$. From these two claims, one can derive $\Box(\mathrm{K}(r \leftrightarrow \mathrm{A}r) \rightarrow (r \leftrightarrow \neg r))$, from which the conclusion follows. (Chalmers 2011, pp 411–12)

The intuitive argument goes like this. Every world is either an r-world or a $\neg r$-world. There is no r-world in which someone knows that $r \leftrightarrow \mathrm{A}r$, because there is no r-world in which someone entertains that $r \leftrightarrow \mathrm{A}r$. One must entertain that $r \leftrightarrow \mathrm{A}r$ in order to know that $r \leftrightarrow \mathrm{A}r$, and one must entertain that q in order to entertain that $r \leftrightarrow \mathrm{A}r$, but the r-worlds are precisely the worlds in which no one entertains that q. At the same time, there is no $\neg r$-world in which someone knows that $r \leftrightarrow \mathrm{A}r$, as $r \leftrightarrow \mathrm{A}r$ is false at $\neg r$-worlds.

I think that Chalmers's argument can be resisted. As before, we should respond in the form of a dilemma. If '$\Diamond$' and '$\Box$' denote metaphysical possibility and necessity, then the argument might be valid and its conclusion might be true, but the paradox is falsely advertised. Its conclusion, then, is that it is metaphysically impossible for someone to know that $r \leftrightarrow \mathrm{A}r$. I accept this conclusion. But the fact that $r \leftrightarrow \mathrm{A}r$ is necessarily unknown does not imply that $r \leftrightarrow \mathrm{A}r$

 © Spencer 2016

is unknowable. There are knowable truths that are necessarily unknown. I gave one example in *Lonely G*, and I think that $r \leftrightarrow \mathrm{A}r$ is another example.

Chalmers's argument has the paradoxical conclusion that it advertises only if '$\Diamond$' and '$\Box$' denote agentive possibility and necessity. But if '$\Diamond$' and '$\Box$' denote agentive possibility and necessity, then premiss (2) is false. Although Ar might imply that it is *metaphysically* necessary that Ar, it does not imply that it is *agentively* necessary that Ar. (Put another way, the fact that no one actually entertains that q does not imply that no one is able to actually entertain that q.) We are supposed to accept (2) as a consequence of a more general principle, namely, $\forall p(\mathrm{A}p \rightarrow \Box \mathrm{A}p)$. But if '$\Box$' denotes agentive necessity, then $\forall p(\mathrm{A}p \rightarrow \Box \mathrm{A}p)$ is the principle of actualized fatalism; it says that agents are able to *actually do* only what they actually do. And just as the principle of fatalism (i.e. $\forall p(p \rightarrow \Box p)$) stands refuted by unexercised abilities, so the principle of actualized fatalism stands refuted by unexercised actualized abilities. If I reach a fork in the road, then I am able to *actually go left* and able to *actually go right*, even though I actually go only one way or the other. In the same sort of way, the fact that I do not actually entertain that q does not imply that I am not able to actually entertain that q; my ability might be unexercised (and hence necessarily unexercised).[46]

We should, in the end, then, reject Chalmers's argument. Those who believe that, for all entertainable p, it is knowable that (p iff actually p) may continue doing so.

But Chalmers's argument is still very interesting, I think. Chalmers has drawn attention to a particularly interesting factive ability—the

[46] Chalmers (2011, pp. 417–18) considers an objection along these lines: 'One … strategy appeals to agentive possibility … holding that a proposition is knowable when it is agentively possible for someone to know it, while denying that agentive possibility entails metaphysical possibility. This position faces an obvious challenge, however, in that the original argument might be reformulated in terms of agentive possibility, and the case for the four key premisses remains strong when read this way.' I respond, as I do in the text, that the case for (2) is not strong when read in terms of agentive possibility.

In personal communication Chalmers has asked me whether we could introduce an actuality-like operator, $\mathcal{A}$, which, still reading '$\Box$' as agentive necessity, makes premiss (2) true. I answer: perhaps we could, and it would be interesting if we could, but introducing this operator does not save the paradox. The widespread belief that all entertainable truths of the form (p iff actually p) are knowable concerns the familiar notion of actuality, which does not imply agentive necessity. If there is an operator, $\mathcal{A}$, which stands to agentive necessity as the familiar notion of actuality stands to metaphysical necessity, then Chalmers has proved that there are entertainable but unknowable truths of the form (p iff $\mathcal{A}p$). But $\mathcal{A}$ is unfamiliar, so the degree to which we should be surprised by this conclusion is unclear.

 © Spencer 2016

ability to know that $r \leftrightarrow Ar$—and proved that this ability is necessarily unexercised. Although necessarily unexercised, this ability is nevertheless widely possessed. Just about anyone who is able to entertain that $r \leftrightarrow Ar$ is able to know that $r \leftrightarrow Ar$. You and I and others like us are able to know that $r \leftrightarrow Ar$, for example—which is interesting. For the fact that you and I and others like us are able to know that $r \leftrightarrow Ar$ implies that, like G, in *Lonely G*, we are able to do something that not a single being in all of modal space manages to do. You and I and others like us literally are able to do the impossible.[47]

References

Adams, Robert Merrihew 1994: *Leibniz: Determinist, Theist, Idealist.* Oxford: Oxford University Press.

Anderson, Alan Ross 1958: 'A Reduction of Deontic Logic to Alethic Modal Logic'. *Mind*, 67, pp. 100–3.

Bennett, Karen 2004: 'Spatio-temporal Coincidence and the Grounding Problem'. *Philosophical Studies*, 118, pp. 339–71.

Berto, Francesco 2013: 'Impossible Worlds'. In Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Winter 2013 Edition).* http://plato.stanford.edu/archives/win2013/entries/impossible-worlds/.

Brand, Myles, and Douglas N. Walton (eds.) 1976: *Action Theory.* Dordrecht: Reidel.

Chalmers, David J. 2011: 'Actuality and Knowability'. *Analysis*, 71, pp. 411–19.

Conee, Earl 1991: 'The Possibility of Power Beyond Possibility'. *Philosophical Perspectives*, 5, pp. 447–73.

Descartes, René 1897: *Oeuvres de Descartes.* Edited by C. Adamand P. Tannery.Paris: Cerf.

——1991: *The Philosophical Writings of Descartes, Volume III: The Correspondence.* Translated by John Cottingham,

[47] Thanks to Selim Berker, David Chalmers, Sam Cumming, Nilanjan Das, Shamik Dasgupta, Katie Elliott, Nate Gadd, Gabriel Greenberg, Ned Hall, Caspar Hare, Liz Harman, Pamela Hieronymi, Wes Holliday, Mark Johnston, Boris Kment, Sarah-Jane Leslie, Eden Lin, John Maier, Matthew McGrath, Calvin Normore, Joe Rachiele, Gideon Rosen, Bernhard Salow, James Shaw, Kieran Setiya, Jonathan Vogel, Ian Wells, and Steve Yablo; to Mind's editor and referees; and to audiences at Princeton (2011, 2012), University of Pittsburgh (2013), MIT (2013), UCLA (2013), University of Missouri (2013), University of Oklahoma (2013), and Harvard (2014).

 © Spencer 2016

Robert Stoothoff, Dugald Murdoch, and Anthony Kenny. Cambridge: Cambridge University Press.

Dummett, Michael 2001: 'Victor's Error'. *Analysis*, 61, pp. 1–2.

——2009: *'Fitch's Paradox of Knowability'*. In Salerno 2009, pp. 51–2.

Eikmeyer, Hans-Jürgen, and Hannes Riemer (eds.) 1981: *Words, Worlds, and Contexts: New Approaches in Word Semantics.* Berlin: de Gruyter.

Eklund, Matti 2008: 'The Picture of Reality as an Amorphous Lump'. In *Sider, Hawthorne and Zimmerman* 2008, pp. 382–96.

Fara, Michael 2008: 'Masked Abilities and Compatibilism'. *Mind*, 117, pp. 843–65.

——2010: 'Knowability and the Capacity to Know'. *Synthese*, 173, pp. 53–73.

Finch, Alicia, and Ted A. Warfield 1998: 'The *Mind* Argument and Libertarianism'. *Mind*, 117, pp. 843–65.

Fine, Kit 1999: 'Things and their Parts'. *Midwest Studies in Philosophy*, 23, pp. 61–74.

Fischer, John Martin 1994: *The Metaphysics of Free Will.* Malden, MA: Blackwell.

——2006a: *My Way: Essays on Moral Responsibility.* New York: Oxford University Press.

——2006b: 'A Framework for Moral Responsibility'. In *Fischer 2006a*, pp. 1–37.

Fitch, Frederic 1963: 'A Logical Analysis of Some Value Concepts'. *Journal of Symbolic Logic*, 28, pp. 135–42.

Frankfurt, Harry G. 1969: 'Alternative Possibilities and Moral Responsibility'. *Journal of Philosophy*, 66, pp. 829–39.

——1977: 'Descartes on the Creation of the Eternal Truths'. *Philosophical Review*, 86, pp. 36–57.

Ginet, Carl 1966: 'Might We Have No Choice?' In *Lehrer 1966*, pp. 87–104.

Hand, Michael 2014: 'Antirealism and Truths Never Known'. *Journal of Philosophy*, 111, pp. 113–34.

Hawthorne, John 2006: *Metaphysical Essays.* Oxford: Oxford University Press.

Hilpinen, Risto (ed.) 1971: *Deontic Logic: Introductory and Systematic Readings.* Dordrecht: Springer.

Holliday, Wesley H. 2012: 'Freedom and the Fixity of the Past'. *Philosophical Review*, 121, pp. 179–207.

 © Spencer 2016

Huemer, Michael 2000: 'Van Inwagen's Consequence Argument'. *Philosophical Review*, 109, pp. 525–44.
Johnston, Mark 2006: 'Hylomorphism'. *Journal of Philosophy*, 103, pp. 652–98.
Kanger, Stig 1957: 'New Foundations for Ethical Theory'. In *Hilpinen 1971*, pp. 36–58.
Kaufman, Dan 2002: 'Descartes's Creation Doctrine and Modality'. *Australasian Journal of Philosophy*, 80, pp. 24–41.
Kearns, Stephen 2011: 'Responsibility for Necessities'. *Philosophical Studies*, 155, pp. 307–24.
Kment, Boris 2010: 'Causation: Determination and Difference-Making'. *Noûs*, 44, pp. 80–111.
——2014: *Modality and Explanatory Reasoning*. Oxford: Oxford University Press.
Kment, Boris MS: 'Freedom, Luck, and the Principle of Sufficient Reason'. Unpublished manuscript.
Kratzer, Angelika 1977: 'What "Must" and "Can" Must and Can Mean'. *Linguistics and Philosophy*, 1, pp. 337–55.
——1981: 'The Notional Category of Modality'. In *Eikmeyere and Rieser 1981*, pp. 38–74.
Lamb, James 1977: 'On a Proof of Incompatibilism'. *Philosophical Review*, 86, pp. 20–35.
Lehrer, Keith (ed.) 1966: *Freedom and Determinism*. New York: Random House.
——1976: ' "Can" in Theory and Practice: A Possible World Analysis'. In *Brand and Walton 1976*, pp. 241–70.
Leslie, Sarah-Jane 2011: 'Essence, Plenitude, and Paradox'. *Philosophical Perspectives*, 25, pp. 277–96.
Lewis, David 1976: 'The Paradoxes of Time Travel'. *American Philosophical Quarterly*, 13, pp. 145–52.
——1981: 'Are We Free to Break the Laws?' *Theoria*, 47, pp. 113–21.
Lin, Martin 2012: 'Rationalism and Necessitarianism'. *Noûs*, 46(3), pp. 418–48.
Lokhorst, Gert-Jan C. 2006: 'Andersonian Deontic Logic, Propositional Quantification, and Mally'. *Notre Dame Journal of Philosophical Logic*, 47, pp. 385–95.
Mares, Edwin J. 1992: 'Andersonian Deontic Logic'. *Theoria*, 58, pp. 3–30.
Mele, Alfred R. 2002: 'Agents' Abilities'. *Noûs*, 37, pp. 447–70.
Moore, G. E. 1912: *Ethics*. London: Williams and Norgate.

 © Spencer 2016

O'Connor, Timothy 2000: *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.
Pike, Nelson 1965: 'Divine Omniscience and Voluntary Action'. *Philosophical Review*, 74, pp. 27–46.
Salerno, Joe (ed.) 2009: *New Essays on the Knowability Paradox*. Oxford: Oxford University Press.
Sider, Theodore, John Hawthorne ,and Dean Zimmerman (eds.) 2008: *Contemporary Debates in Metaphysics*. Malden, MA: Blackwell.
Skow, Bradford MS: 'Notes on the Grandfather Paradox'.
Stalanker, Robert 1996: 'Impossibilities'. *Philosophical Topics*, 24, pp. 193–204.
van Inwagen, Peter 1978: 'Ability and Responsibility'. *Philosophical Review*, 87, pp. 201–24.
van Inwagen 1983: *An Essay on Free Will*. Oxford: Oxford University Press.
Vetter, Barbara 2013: ' "Can" without Possible Worlds: Semantics for Anti-Humeans'. *Philosophers' Imprint*, 13(16), pp. 1–27.
——2014: 'Dispositions without Conditionals'. *Mind*, 123, pp. 129–56.
Vihvelin, Kadri 1996: 'What Time Travelers Cannot Do'. *Philosophical Studies*, 81, pp. 315–30.
——2011: 'Arguments for Incompatibilism'. In Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*. http://plato.stanford.edu/archives/spr2011/entries/incompatibilism-arguments/.
——2013: *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oxford: Oxford University Press.
Vranas, Peter B. M. 2010: 'What Time Travelers May Be Able to Do'. *Philosophical Studies*, 150, pp. 115–21.
Wolf, Susan 1990: *Freedom Within Reason*. Oxford: Oxford University Press.
Yablo, Stephen 1987: 'Identity, Essence, and Indiscernibility'. *Journal of Philosophy*, 84, pp. 293–314.

 © Spencer 2016

Everything, and Then Some

Stephan Krämer
University of Hamburg
stephan.kraemer@uni-hamburg.de

On its intended interpretation, logical, mathematical and metaphysical discourse sometimes seems to involve absolutely unrestricted quantification. Yet our standard semantic theories do not allow for interpretations of a language as expressing absolute generality. A prominent strategy for defending absolute generality, influentially proposed by Timothy Williamson in his paper 'Everything' (2003), avails itself of a hierarchy of quantifiers of ever increasing orders to develop non-standard semantic theories that do provide for such interpretations. However, as emphasized by Øystein Linnebo and Agustín Rayo (2012), there is pressure on this view to extend the quantificational hierarchy beyond the finite level, and, relatedly, to allow for a cumulative conception of the hierarchy. In his recent book, *Modal Logic as Metaphysics* (2013), Williamson yields to that pressure. I show that the emerging cumulative higher-orderist theory has implications of a strongly generality-relativist flavour, and consequently undermines much of the spirit of generality absolutism that Williamson set out to defend.

1. Introduction

In ordinary discourse, most of our quantifications are restricted to a set of contextually relevant objects. I say 'There is no beer', meaning not that there is absolutely no beer in the entire universe, but that there is no beer in my fridge, and thus no contextually relevant beer. In logical, metaphysical and mathematical discourse, in contrast, we often seem to generalize without any such restrictions. In typical utterances of 'Nothing has contradictory properties', 'Everything is self-identical', 'Everything is either abstract or concrete', or 'Nothing is a member of the empty set', it seems, absolutely nothing is excluded as contextually irrelevant.

However, the appearance that absolute generality can thus be expressed comes under pressure from a number of theoretical considerations.[1] The one that is most important for our purposes is that in our most successful and best understood semantic theories,

[1] For an excellent overview of the debate, including a discussion of other arguments against the possibility of absolute generality, see Rayo and Uzquiano (2006a), as well as Florio (2014).

Mind, Vol. 126 . 502 . April 2017
© Krämer 2016
doi:10.1093/mind/fzv187 Advance Access publication 30 September 2016

quantification is always interpreted with respect to some set that constitutes the presumed domain of discourse. Since there is no universal set, these semantic theories do not allow for absolute generality. Some philosophers hold that this apparent limitation of such theories cannot be overcome in a satisfactory manner, and have therefore embraced *(Generality) Relativism*, the view that, initial appearances notwithstanding, there can be no such thing as absolutely general discourse.[2]

In his paper 'Everything' (2003), Timothy Williamson mounts a forceful defence of the opposition to Relativism, that is, *(Generality) Absolutism*. As part of this defence, he proposes an alternative kind of semantics in which object-language quantifiers need not be interpreted as restricted to a set.[3] The crucial move that enables him to avoid this limitation is the employment of *higher-order* quantification in the metalanguage.[4] As Williamson notes, the obvious generalization of his proposal requires the use of at least the whole finite hierarchy of orders of quantification. Moreover, as Øystein Linnebo and Agustín Rayo (2012) have shown, given a number of plausible further assumptions, we need to countenance even quantification of *transfinite* and *cumulative* orders. And when Williamson returns to the issue in his *Modal Logic as Metaphysics* (2013), he explicitly avails himself of such quantifiers.

The purpose of this paper is to show that the move to transfinite, cumulative orders of quantification undermines much of the spirit, if not the letter, of the Absolutist picture that Williamson wishes to defend. Moreover, since the relevant components of Williamson's view are difficult or impossible to reject given his basic approach of using higher-order resources to defend absolute generality, the

[2] Sometimes in the debate over absolute generality a distinction is made between a *metaphysical* question—roughly, whether there is an all-inclusive domain of discourse—and a linguistic or *availability* question—roughly, whether we could quantify over such a domain, if it existed (see Rayo and Uzquiano 2006a, p. 2, where they credit Kit Fine with having first emphasized this point). The distinction, if it can be made, is not of particular importance for our present purposes. As I understand the distinction, our discussion concerns the availability question throughout.

[3] It may be objected that familiar truth-theoretic semantics formulated in the Davidson-Tarski style have no need to interpret object-language quantifiers as restricted to a set. That is, of course, correct; the difficulty arises when we try to formulate a theory that specifies the truth conditions of the object-language sentences for arbitrary interpretations of the language. We shall come back to this point shortly.

[4] The same kind of move has been employed for a similar objective by George Boolos (1985) and, further developing Boolos's suggestions, by Rayo and Uzquiano (1999).

 © Krämer 2016

argument also presents a severe challenge to any view following this general approach. I first present the current state of the debate. §2 describes the problem that Absolutists face when trying to formulate an adequate semantics for absolutely general quantification. §3 presents Williamson's higher-orderist solution to the problem, and then explains both the notion of quantification at transfinite and cumulative orders and why Williamson avails himself of such quantifiers. §4 develops my argument that the cumulative higher-order picture has implications strongly reminiscent of Relativism. I first state the argument in informal terms and indicate the obstacles to formalizing it in the non-cumulative higher-orderist's canonical language. I then show how the introduction of cumulative resources allows us to overcome these obstacles. The crucial bit of cumulative ideology that we need is a higher-order, cumulative analogue of the notion of identity. That notion is discussed in more detail in §5. Finally, I ask to what extent the problems Williamson raises for Relativism have counterparts that apply to his own cumulative higher-orderist version of Absolutism. §6 points out that cumulative higher-orderism faces a similar difficulty to standard Relativism concerning the adequate formulation of its apparent Relativist commitments. §7 turns to the criticism of standard Relativism that its restricted interpretations often seem to constitute weird misinterpretations of relevant object-language discourse. I show that this criticism also applies on the cumulative higher-orderist picture, though in a somewhat less dramatic form. I conclude in §8 that Williamson's cumulative higher-orderist Absolutism is a lot closer to Relativism than it first appears.

2. The semantic challenge for absolutism

A peculiar feature of the dispute over absolute generality is that it is not very easy to say just what it takes to be an Absolutist, or a Relativist, who is worthy of the name.[5] Take Absolutism. It is not enough to say 'It is possible to quantify over absolutely everything', or even

(1) I am now quantifying over absolutely everything.

The reason is that *everyone* can say that, since any utterance of (1) will express a truth. For whatever the quantifier phrase 'absolutely everything' ranges over in such an utterance, what the speaker then claims

[5] On this issue, see Williamson (2003, §V), Florio (2014, p. 2), Rayo and Uzquiano (2006a, pp. 2f.), and the references given there.

 © Krämer 2016

is merely that they are quantifying over everything in that range, which cannot fail to be true. However, if the quantifier phrase ranges over only a restricted domain, the claim made by the speaker, though true, is irrelevant to the spirit of Absolutism, if perhaps not to its letter. It is only if the quantifier phrase ranges over absolutely everything that the speaker makes a claim that is relevant to the debate, such that accepting it makes one an Absolutist proper.

If the speaker is to count as an Absolutist, therefore, we shall require them to back up their utterance of (1) by insisting, in a metalinguistic utterance, on an interpretation of their utterance on which it generalizes over everything. And while the speaker remains firmly on the safe side with their utterance of (1), once they back it up in this way in the metalanguage, as Williamson shows, they face a real threat of contradiction. First, let us shift our focus slightly by considering formal languages instead of utterances of English sentences like (1). Presumably, if it can be consistently maintained of an utterance of (1) that it generalizes over everything, then it is possible to specify a formal language of which it can be consistently maintained that its quantifiers range over everything. We shall therefore assume that the Absolutist commits to the following claim:

> (GA) It is possible to specify a formal language containing quantifiers that, from the point of view of a suitable metalanguage, range over absolutely everything.

A language $\mathcal{L}'$ is a suitable metalanguage for a formal language $\mathcal{L}$ if it is possible to develop an adequate semantic theory for $\mathcal{L}$ in $\mathcal{L}'$.

A second assumption we shall make has to do with the *kind* of semantic theorizing about the object language that we want to be possible in our metalanguage. We shall assume that for any formal language it is possible to construct what we may call a *generalized semantics* for that language.[6] A generalized semantics for a language is a semantic theory which provides an inductive characterization of the truth conditions of the object-language sentences, relative to arbitrary ways of interpreting (at least) their non-logical constants. (A generalized semantics is therefore exactly the kind of theory we require as the basis of a broadly model-theoretic theory of logical conse-

[6] The way I set up the dispute here is indebted to Linnebo (2006) and Linnebo and Rayo (2012), from which I have also borrowed my terminology.

 © Krämer 2016

quence.) Our second assumption can thus be described as a principle of *Semantic Optimism*:

> (SO) For any formal language $\mathcal{L}$, it is in principle possible to construct a theory adequately specifying the truth conditions of sentences of $\mathcal{L}$ relative to arbitrary ways of interpreting $\mathcal{L}$'s non-logical constants.

This principle is by no means trivial. Still, in the debate about absolute generality, it is standardly assumed that the principle enjoys a high degree of initial plausibility (see, for example, Linnebo 2006, p. 150; Linnebo and Rayo 2012, pp. 276f.). For the purposes of this paper, I shall therefore take it for granted. The challenge for the Absolutist is then to specify a formal language, and to formulate a generalized semantics for that language, so that from the point of view of the metalanguage, the quantifiers of the object language range over absolutely everything. Unfortunately, there is a powerful argument purporting to show that this cannot be done.[7]

Suppose that the Absolutist has described a formal language $\mathcal{L}_1$ whose (first-order) quantifiers they wish to maintain express absolute generality. The Absolutist must now formulate a generalized semantics for $\mathcal{L}_1$ in a suitable metalanguage. A generalized semantics is supposed to specify the truth conditions of the sentences of $\mathcal{L}_1$ relative to arbitrary ways of interpreting (the non-logical constants of) $\mathcal{L}_1$. It therefore needs to generalize over (things that model) ways of interpreting $\mathcal{L}_1$. For simplicity, assume that '*i*' is a metalanguage variable that ranges over whatever the semantics uses to model ways of interpreting $\mathcal{L}_1$.

Now take some monadic predicate *P* of $\mathcal{L}_1$. It seems very plausible that we can in principle use any contentful monadic predicate of the metalanguage to interpret *P*. If so, then any instance of the following *Comprehension Schema for Predicate Interpretations*, obtained by replacing '*F*' with such a metalanguage predicate, should be true.

> (CPI Informal) Under at least one way of interpreting the non-logical constants of $\mathcal{L}_1$, *P* applies to all and only the *F*s in the range of $\mathcal{L}_1$'s quantifiers.

The semantics should then imply a suitable regimentation of (CPI Informal) in the metalanguage. For brevity, we add copies '$\forall_o$' and '$\exists_o$' of $\mathcal{L}_1$'s first-order quantifiers to the metalanguage. We may then

[7] Except for minor details, the argument to follow is Williamson's (2003, pp. 425ff.).

 © Krämer 2016

formulate a suitable regimentation of (CPI Informal) as follows, using 'applies_i' to abbreviate 'applies under interpretation i':[8]

(CPI Formal 1) $\exists i \forall_o x$ (P applies_i to $x \leftrightarrow Fx$)

The Absolutist also has to include in their theory an expression of the claim that, from the point of view of the semantics, $\mathcal{L}_1$'s quantifiers range over absolutely everything. An obvious way to formalize that claim is:

(GA_1) $\forall y \exists_o x\ x = y$

However, these assumptions jointly entail a contradiction. For consider the following instance of (CPI Formal 1), obtained by substituting '$\neg P$ applies_x to x' for 'Fx':[9]

(2) $\exists i \forall_o x$ (P applies_i to x iff $\neg P$ applies_x to x)

Assuming that some interpretation i' verifies (2), we have:

(3) $\forall_o x$ (P $\text{applies}_{i'}$ to x iff $\neg P$ applies_x to x)

Now crucially, by (GA_1), the range of '$\forall_o x$' comprises everything. So, in particular, it includes the interpretation i'. We may therefore instantiate (3) with i' to obtain:

(4) P $\text{applies}_{i'}$ to i' iff $\neg P$ $\text{applies}_{i'}$ to i'

But (4) is equivalent in classical logic to an explicit contradiction.

The Relativist maintains that it is the Absolutist's claim (GA_1) which is to blame for the contradiction, and concludes that there is more than is dreamt of by $\mathcal{L}_1$'s quantifiers:

(MORE_1) $\exists y \forall_o x\ x \neq y$

To be an Absolutist, in contrast, one must keep (GA_1), and thus find some other way out. The next section explains how Williamson proposes to do that.

3. Higher-orderist absolutism

Our above regimentation of (CPI Informal) by (CPI Formal 1) embodies a tacit assumption to the effect that the informal talk of ways of interpreting an object-language predicate is to be understood in *first-order* quantificational terms. After all, (CPI Formal 1) uses a first-order quantifier '$\exists i$' presumed to range over objects of some sort that are

[8] The variable 'i' is not allowed to occur free in a substituend for 'F'.

[9] Note that since 'i' in 'applies_i' was assumed to be an ordinary first-order variable, we can legitimately put the first-order variable 'x' in its place.

 © Krämer 2016

identified with, or taken to represent, ways of interpreting object-language predicates. Williamson suggests that this tacit assumption is mistaken; it is more plausible, according to him, to represent ways of interpreting predicates by means of a *second-order* variable (2003, pp. 452ff.). In effect, his proposal has us replace (CPI Formal 1) with a second-order analogue, in which the first-order variable '*i*' and the quantifier binding it have been replaced by a second-order variable and quantifier:[10]

(CPI Formal 2) $\exists I \forall_o x$ (P applies$_I$ to $x \leftrightarrow Fx$)

The semantic predicate 'applies$_I$' may be thought of as defined by:

(Df. applies$_I$) P applies$_I$ to $x \leftrightarrow I(P, x)$

Note that from the standard (full) axiom scheme of comprehension for second-order logic, we obtain every instance of the schema

(5) $\exists I \forall_o x$ ($I(P, x) \leftrightarrow Fx$)

in which the expression replacing 'F' does not contain 'I' free. This in turn guarantees the validity of (CPI Formal 2).

At least at first glance, it seems to me, this approach to the semantic challenge for Absolutism—call it the *higher-orderist* approach—has a lot to be said for it. What the Russell-paradoxical argument from (2) to (4) shows, we might say, is that there are always strictly more ways to interpret a predicate with respect to a domain of objects than there are objects in that domain, and thus ways to interpret a first-order variable with respect to that domain.[11] As a result, not every way to interpret a predicate with respect to a given domain can be represented by an object in that domain. The Relativist concludes from this that there must always be objects outside any domain with respect to which a predicate can be interpreted. The higher-orderist approach, in contrast, concludes that it is a mistake to try to represent ways of interpreting a predicate by the values of first-order variables. Instead, the thought goes, we should use second-order variables for that purpose. For crucially, there is no obstacle to holding that there are always at most as many ways to interpret a predicate with respect to a domain of objects as there are ways to interpret a predicate variable with respect to that domain.

[10] Similarly as for (CPI Formal 1), instantiation of 'F' with a predicate in which 'I' is free is not allowed.

[11] The talk of domains is not to be taken too literally. There is no need to assume that the objects that a predicate is interpreted as true or false of are members of some further object that we call a domain.

 © Krämer 2016

It is important to note, however, that the proposal depends on a very specific view—call it *higher-orderism*—of second- and higher-order quantification. Before stating the view, let me clarify my talk of quantifiers of first, second and higher orders. For present purposes, that classification is to be thought of in *syntactic* terms. On this understanding, what makes the familiar quantifiers '$\exists x$' and '$\forall x$' of $\mathcal{L}_1$ first-order is that they bind variables that stand in the syntactic position of singular terms. What we shall call *second-order* quantifiers are then quantifiers that bind variables standing in the syntactic position of expressions that form sentences when combined with one or more singular terms as their arguments, that is, ordinary *predicates*. A *third-order* quantifier, by analogy, is a quantifier binding variables that take the position of expressions forming sentences when combined with ordinary predicates as their arguments. The hierarchy extends in the obvious way to quantifiers and variables of order n for any finite n.

Higher-orderism can now be defined as the conjunction of the following three theses concerning this syntactic hierarchy. (1) Quantification of any finite order is a legitimate linguistic device. (2) Quantifications of a given order are not in general paraphrasable by quantifications of a lower order. (3) For any finite n, adequate semantic clauses for nth-order quantifications themselves employ nth-order quantifiers of the metalanguage in the way standard clauses for first-order quantifiers employ first-order quantifiers of the metalanguage.[12]

To get an idea of why the Williamsonian proposal requires the full strength of higher-orderism, note first that if second-order quantification were in general paraphrasable by first-order quantification, then in particular (CPI Formal 2) could be paraphrased in first-order terms. But such a paraphrase would reintroduce the inconsistency engendered by (CPI Formal 1). Moreover, given (SO), we can develop a generalized semantics for our second-order metalanguage. If we were to do that in a first-order metalanguage, perhaps by construing second-order quantifiers as ranging over properties, we should again run into a version of Russell's Paradox. Indeed, by a higher-order analogue of the above Russell-style argument, it can be shown that a generalized semantics for a second-order language cannot be given even in a second-order language: we need to use third-order

[12] I am not claiming that the three theses are independent, but it is a non-trivial question what entailment relations may obtain between them, and so it seems best to stay neutral with respect to that question in characterizing higher-orderism.

 © Krämer 2016

resources.[13] So Williamson's higher-orderist approach depends on the legitimacy and irreducibility of quantification of every finite order.

The irreducibilism embodied in higher-orderism has important implications for how we can read second- and higher-order quantifications in natural language. For example, at first glance the second-order quantification '$\exists X\ \forall x\ \neg Xx$' might naturally be read as 'some property is had by no object'. However, although often useful and appropriate for heuristic purposes, such a reading cannot be considered strictly adequate on the higher-orderist view. The reason is that the English quantifier here used to interpret the formal higher-order quantifier is itself *first-order*. This can be seen from the fact that if we ask for a witnessing instance of the quantification, grammar demands that the answer consist in a singular noun phrase like 'the property of being self-distinct' rather than a predicate.

It is controversial whether one can translate second-order quantification into natural language in a way that fits higher-orderism.[14] For our purposes, it does not matter too much, for even if this is possible, it seems clear that natural languages do not provide us with the resources needed to appropriately translate quantifiers of arbitrarily high finite orders. So higher-orderists must hold that we can, at least in principle, somehow come to understand third-, fourth- and higher-order quantification without the benefit of a translation into vocabulary that we independently understand.[15] In what follows, to avoid excessive formalism, and for heuristic purposes, we shall make use of various natural language constructions to approximate the higher-orderists' intended interpretation of their vocabulary, bearing in mind, though, that these may occasionally yield a slightly misleading picture of the higher-orderist view. We shall grant, moreover, both the truth of higher-orderism and the claim that higher-orderism provides sufficient expressive resources to formulate a generalized semantics for any language of finite order, that is, such that for some finite n, it contains no quantifiers of an order higher than n.

[13] Strictly speaking, the situation is slightly more complicated; it is examined in detail in Rayo (2006). The essential point remains, however. Semantic optimism forces the higher-orderist up the hierarchy of metalanguages, and if they are to permit the formulation of a generalized semantics in accordance with the higher-orderist approach, the metalanguages must include quantificational devices of ever increasing finite orders.

[14] George Boolos (1984, 1985) famously proposed a translation using English plural quantification; an alternative, more predicational reading was first suggested by Arthur Prior (1971, ch. 3) and recently developed in more detail by Rayo and Yablo (2001).

[15] See Williamson (2003, pp. 457ff.), Linnebo (2006, pp. 152ff.).

 © Krämer 2016

So far, I have merely described the explicit key commitments of the higher-orderist defence of absolute generality as described in, for example, Williamson (2003) and Rayo (2006). We now turn to two crucial further claims that Williamson endorses, implicitly at least, in his recent book (2013). The *first* claim is that in addition to the languages of finite order that we have already canvassed, there is also a legitimate language that contains quantifiers of *every* finite order. With respect to the ordering of languages of higher and higher finite orders, this language would occupy the level of the first limit ordinal ω, so we may call it $\mathcal{L}_\omega$. Williamson uses a language like this as his favoured background language for metaphysical theorizing, and so is obviously committed to considering such a language legitimate.[16]

The question whether the higher-orderist defence of absolute generality on its own is committed to this claim is more difficult to answer. Certainly, the legitimacy of $\mathcal{L}_\omega$ does not follow *logically* from anything the higher-orderist has said so far. Nevertheless, it would seem prima facie quite implausible to disallow it.[17] For every bit of vocabulary we find in $\mathcal{L}_\omega$ has already been deemed legitimate, since it is also found in some 'successor' language in the higher-orderist's hierarchy. And it is hard to see how pooling all these individually coherent and legitimate linguistic resources together into a single language could somehow fail to produce an equally coherent and legitimate language. The burden of proof would therefore seem to lie with anyone wishing to deny the legitimacy of the limit language $\mathcal{L}_\omega$.

If $\mathcal{L}_\omega$ is allowed, then by (SO), it is possible to give a generalized semantics for it. The *second* crucial claim of Williamson's is that such a semantics can be stated using quantification of *transfinite* orders. Note first that higher-orderism, as defined above, does not by itself provide sufficient resources to formulate a semantics for $\mathcal{L}_\omega$ in accordance with the higher-orderist approach. For on this approach, to interpret expressions of order n, we need to make use of an interpretation variable of at least order n. But every kind of variable higher-orderism provides us with belongs to some finite order n. And for every finite order n, since $\mathcal{L}_\omega$ contains quantifiers and variables of every finite order, it contains variables of order $n + 1$. So no interpretation variable of a finite order can be used to interpret all the expressions of $\mathcal{L}_\omega$.

[16] The modal higher-order language ML_P proposed by Williamson (2013, ch. 5) contains quantifiers of every finite order, and is thus relevantly like $\mathcal{L}_\omega$.

[17] On this point, see also Linnebo and Rayo (2012, pp. 275f.) and Rayo (2006, pp. 246ff.).

 © Krämer 2016

What Williamson does, therefore, in specifying a semantics for a limit language, is avail himself of predicates, and quantifiable predicate variables, of *transfinite* orders.[18]

Aside from their transfinite character, Williamson's predicates and predicate variables of transfinite order have a further striking feature that sets them apart from any of the expressions we have encountered so far. That feature concerns what expressions they accept as arguments. Usually, a predicate or predicate variable accepts as arguments expressions of the next lower order: a first-order predicate accepts singular terms, a second-order predicate accepts first-order predicates, and so on. This rule cannot extend to our new predicates, though. For these are of order ω, and there is no next lower order to ω. Instead, these predicates accept expressions of *any* lower—that is finite—order in their argument place. I shall therefore describe these new predicates as *syntactically cumulative*. It turns out that a syntactically cumulative ω-order predicate variable is just what is needed, and just what Williamson uses, for an interpretation variable in a generalized semantics for $\mathcal{L}_\omega$.[19]

Let me stress that for the purposes of this paper, the admission of *cumulative* resources is the important point. The admission of $\mathcal{L}_\omega$ and transfinite orders of quantification matter only in virtue of their bearing on this point. And while the above consideration provides perhaps the most principled and compelling case for the legitimacy of cumulativity, it should be noted that even independently of limit languages and transfinite orders of quantification, higher-orderists may have reason to be sympathetic to this claim, for since they allow quantification of every finite order, they are already committed to a fairly liberal standard for admissible linguistic devices. And from a logico-mathematical point of view, cumulative resources make perfect sense

[18] See Williamson (2013, pp. 236ff.). I do not know whether the move to transfinite orders is strictly the *only* way to give a generalized semantics for a language like $\mathcal{L}_\omega$, as seems to be suggested by the discussion in Linnebo and Rayo (2012, p. 275, and appendix B). For all I know, it might also be possible to use separate interpretation variables for every order, which could then themselves all be of finite order. However, we would then have to relativize satisfaction to infinitely many parameters, so we should have to introduce predicates with infinitely many argument places, as well as quantifiers binding an infinite set of variables. I do not know what a semantics that is adequate for this kind of infinitary language would have to look like. However, I cannot entirely rule out that there might be a coherent version of higher-orderism that allows $\mathcal{L}_\omega$ but no transfinite orders of quantification.

[19] The construction is given in Williamson (2013, pp. 236ff.). A more detailed presentation and discussion of similar constructions is given in the appendices of Linnebo and Rayo (2012).

 © Krämer 2016

and have a principled and well-behaved logic.[20] So it is not at all clear that there is a defensible standard of intelligibility that could serve to rule out cumulative devices without at the same time ruling out quantifiers of very high finite orders.

4. Cumulative higher-orderism and the spirit of absolutism

Let us return from the dizzy heights of ω-order quantification to the first and simplest stage in the higher-orderist's hierarchy of languages and semantic theories: the generalized semantics for the first-order language $\mathcal{L}_1$, formulated in a second-order metalanguage. Recall the sentence that started us off on the whole higher-order journey, our formalization of the claim that $\mathcal{L}_1$'s quantifiers are, from the point of view of the semantics for $\mathcal{L}_1$, absolutely general:

(GA_1) $\forall y \exists_o x\ x = y$

This section develops an argument that according to cumulative higher-orderism, even though the higher-order semantics for $\mathcal{L}_1$ includes (GA_1), it does not make the quantifiers of $\mathcal{L}_1$ absolutely general. More precisely, I argue that for a cumulative higher-orderist, there is a good sense in which the following holds:

> (MORE) From the point of view of the higher-order semantics for $\mathcal{L}_1$, there is more than is quantified over in $\mathcal{L}_1$.

I shall begin by sketching the argument in informal terms.

In the setting of the original, first-order semantics for $\mathcal{L}_1$, (GA_1) constitutes an adequate formalization of the claim that $\mathcal{L}_1$'s quantifiers are absolutely general. For the quantifier '$\forall y$' ranges over absolutely every bit of reality that is countenanced in that semantics, and so (GA_1) says of absolutely everything countenanced in the semantics that it is in the reach of the quantifiers of $\mathcal{L}_1$. However, that situation changes when the higher-orderist proceeds to extend the language of the semantics by second-order quantifiers. Since the higher-orderist insists that these are in no way reducible to first-order ones, we have to see them as concerned with new bits of reality that are not in the range of the semantics' first-order quantifiers.[21] But then '$\forall y$' no longer

[20] On this point, see also Linnebo and Rayo (2012, p. 278). For an in-depth discussion of the logico-mathematical properties of cumulative higher-order logic, see Degen and Johannsen (2000).

[21] Higher-orderists may well consider my use of the phrase 'new bits of reality' misleading and inappropriate, so I want to emphasize that I am merely offering an informal sketch of an argument to be developed with formal rigour below. (Thanks here to an anonymous referee.)

 © Krämer 2016

ranges over absolutely every bit of reality that is countenanced in that semantics. Consequently, (GA_1) is no longer a plausible formalization of the absolute generality of $\mathcal{L}_1$'s quantifiers. Even on the higher-orderist picture, there is *more* than is dreamt of by $\mathcal{L}_1$'s quantifiers—what sets the picture apart from the Relativist's is only that what is more is not in the range of the semantics' first-order quantifiers but their second-order cousins. So the higher-orderist version of Absolutism is not really worthy of the name. In subscribing to (GA_1), it preserves the letter of Absolutism, but in implying (MORE), it gives up on its spirit.

Cast as it is in informal terms, the objection so far inspires limited confidence. After all, the higher-orderist has warned us that informal, natural-language approximations of their higher-order quantifications can be misleading. Can we put the informal objection on a more rigorous footing by reproducing it in a formal setting congenial to higher-orderism? More specifically, can we find a plausible formalization of (MORE) which is a consequence of the higher-orderist semantics?

Here is a somewhat flat-footed argument that we cannot. To say, as the objection alleges, that there is more, on the higher-orderist view, than is included in the range of $\mathcal{L}_1$'s quantifiers, we should have to say that there is something which is *distinct* from everything in that range. The condition of being distinct from everything in that range is expressed by

$(\neq_1)\ \forall_o x \ldots \neq x$

Now, to say that there is something satisfying that condition, one has to put a variable into the gap of ($\neq_1$), and bind it by an existential quantifier. If the result is to be well-formed, however, we can only put a *first-order* variable into the gap of ($\neq_1$). But the putative extra bits of reality countenanced by the higher-orderist are supposed to be introduced only by second-order quantifiers. Such a quantifier, however, cannot bind the first-order variable in the gap of ($\neq_1$). So there is no sense in which, according to higher-orderism, there is something *more* than is ranged over by the quantifiers in $\mathcal{L}_1$. Any attempt to even formulate that claim in the higher-orderist's canonical language produces an ill-formed string. In that language, as Williamson puts it, 'quantification into predicate position is simply incommensurable with quantification into name position; the former presents no coherent threat to the absolute generality of the latter' (Williamson 2003, p. 458).

 © Krämer 2016

As I have stated it, this line of reasoning depends on the following assumption:

(DIST) A regimentation of (MORE) contradicts the spirit of Absolutism only if it uses ($\neq_1$) to express distinctness from anything in the range of the object language quantifier.

I shall argue that this assumption is implausible, and that once we are allowed to make use of *cumulative* higher-order resources, second-order quantification ceases to be incommensurable with first-order quantification, and does present a serious threat to the latter's absolute generality. In a first step, I show that (DIST) should be rejected even independently of any issues to do with cumulativity. In a second step, I present a number of relatively modest ways to extend the higher-orderist's second-order metalanguage with cumulative vocabulary, and show that in this extended cumulative higher-orderist setting, we can formulate and prove a well-formed regimentation of (MORE) that clashes with the spirit of Absolutism.

It will help if I first set up a system of grammatical types that allows us to describe the syntax of cumulative as well as non-cumulative expressions. For ease of comparison of the resources I employ to those used by Williamson, I base my system on the one he uses (2013, p. 221). It has just one basic type e, which is the type of singular terms. Then whenever $t_1, \ldots, t_n$ are types, $\langle t_1, \ldots, t_n \rangle$ is the derived (functional) type of expressions that form sentences when combined with n further expressions of types $t_1, \ldots, t_n$, respectively. As a limiting case, we allow $\langle\,\rangle$ as the type of sentences, that is, expressions forming sentences when combined with zero further expressions. We also add a category of cumulative types: whenever $t_1, t_2, \ldots$ are types, $[t_1, t_2, \ldots]$ is the cumulative type including all expressions belonging to any of $t_1, t_2, \ldots$[22] We write '$\langle e* \rangle$' to abbreviate the infinite string '$\langle e \rangle, \langle e, e \rangle, \langle e, e, e \rangle, \ldots$' and similarly for other types. So $[\langle e* \rangle]$ is the cumulative type including every (first-order) predicate, of whatever adicity. Since cumulative types are types, the recursive clause for functional types now also yields new functional types. There is, for example, the functional type $\langle e, [e, \langle e* \rangle] \rangle$ of expressions forming

[22] For the metalanguage of his limit language, Williamson adds only one cumulative type, namely, a cumulative infinite limit type, λ, that comprises exactly the expressions belonging to any finite type (2013, p. 221). For my purposes, it is simpler to use only smaller cumulative types at finite orders. There should be no objection to this. If we can form a cumulative infinite limit type like λ and use expressions of types derived from it, surely we can also form a cumulative type comprising, say, only names and first-order predicates, and use expressions of types derived from it.

 © Krämer 2016

sentences when combined with a name as their first argument and an expression of type $[e, \langle e* \rangle]$ as their second argument. Since there are both names and predicates among the latter, such expressions accept, in the same argument place, both names and predicates.

We use this system of types to describe, in the first step, an ordinary, non-cumulative second-order language $\mathcal{L}_2$. It includes, for each type τ among e, $\langle e* \rangle$, a countably infinite stock of constants and variables. We write the constants using lower-case letters from the beginning of the alphabet, augmented with subscripts as required, and marking their type by a superscript. We adopt the same convention for variables, except for choosing letters from the end of the alphabet. $\mathcal{L}_2$ also includes the identity predicate '$=^{\langle e, e \rangle}$', the usual quantifier symbols, connectives, and parentheses. Type-superscripts may be omitted if there is no risk of ambiguity, and parentheses may be omitted or added according as readability is improved.

$\mathcal{L}_2$ does not include any cumulative devices, so we may call it a *pure* second-order language. In all relevant ways, it is exactly the kind of language the higher-orderist needs to formulate his generalized semantics for the first-order language $\mathcal{L}_1$. If we extend it by copies '$\forall_o x^e$' and '$\exists_o x^e$' of the quantifiers in $\mathcal{L}_1$, we can formulate this version of (GA_1):

$(\mathrm{GA}_1{}^*)$ $\forall y^e \exists_o x^e \; x^e = y^e$

Given the requisite amount of syntax and set theory, we can go on to formulate in $\mathcal{L}_2$ a higher-orderist generalized semantics for $\mathcal{L}_1$ that includes $(\mathrm{GA}_1{}^*)$.

I shall now argue that, independently of the admissibility of cumulative expressions, (DIST) should be rejected. The reason is that although the identity predicate '=' itself does not apply at the level of second- and higher-order quantification, higher-order analogues of that predicate do apply. Using these higher-order analogues of the identity predicate, we can construct higher-order counterparts of $(\neq_1)$ and thereby obtain higher-order analogues of (MORE_1). I maintain that these are similar enough to (MORE_1) that they should be taken to contradict the spirit of Absolutism.

We wish to extend $\mathcal{L}_2$ by non-cumulative higher-order cousins of the identity predicate connecting predicates and predicate variables of a given adicity. These expressions should then belong to the types $\langle\langle e \rangle, \langle e \rangle\rangle$, $\langle\langle e, e \rangle, \langle e, e \rangle\rangle$, etc. We shall officially write them as '$=^{\langle\langle e \rangle, \langle e \rangle\rangle}$', '$=^{\langle\langle e, e \rangle, \langle e, e \rangle\rangle}$', etc., but often let context fix the type. What should we take these predicates to mean? For present purposes,

 © Krämer 2016

we can think of them in one of three ways. Firstly, we can take them as primitive expressions subject to inference rules analogous to the standard rules for identity. Where 'a' and 'b' belong to some type τ among $\langle e\rangle$, $\langle e, e\rangle$, ..., and $\Phi^{a/b}$ is the result of replacing zero or more occurrences of 'b' in Φ by 'a':

$(=_1) \vdash a =^{\langle \tau, \tau\rangle} a$

$(=_2) \vdash a =^{\langle \tau, \tau\rangle} b \rightarrow (\Phi \rightarrow \Phi^{a/b})$

Secondly, we can extend $\mathcal{L}_2$ by third-order variables of type $\langle\langle e\rangle\rangle$ ($\langle\langle e, e\rangle\rangle$, ...) and quantifiers binding them. The expression '$x^{\langle e\rangle} = y^{\langle e\rangle}$' can then be taken to abbreviate the corresponding indiscernibility condition: $\forall z^{\langle\langle e\rangle\rangle}(z^{\langle\langle e\rangle\rangle}(x^{\langle e\rangle}) \leftrightarrow z^{\langle\langle e\rangle\rangle}(y^{\langle e\rangle}))$, and similarly for the other types. Thirdly, since $\mathcal{L}_2$ is an extensional language, co-extensiveness implies indiscernibility, so we can take the predicates simply to abbreviate the relevant co-extensiveness condition, so that, for example, '$x^{\langle e\rangle} = y^{\langle e\rangle}$' abbreviates '$\forall z^e(x^{\langle e\rangle}(z^e) \leftrightarrow y^{\langle e\rangle}(z^e))$'. Under both the second and third option, the rules $(=_1)$ and $(=_2)$ are derivable.

Now suppose we include in $\mathcal{L}_2$ a copy '$\forall_o x^{\langle e\rangle}$' of a monadic second-order quantifier of some second-order language. We may wonder whether, from the perspective of $\mathcal{L}_2$, that quantifier expresses an unrestricted, absolute form of monadic second-order generality. Roughly speaking, that is, we may wonder whether there is something in the range of $\mathcal{L}_2$'s monadic second-order quantifier that is distinct from anything in the range of '$\forall_o x^{\langle e\rangle}$'. The informal talk of being distinct from anything in that range here can be regimented by means of '$=^{\langle\langle e\rangle, \langle e\rangle\rangle}$':

$(\neq_2) \forall_o x^{\langle e\rangle} \ldots \neq^{\langle\langle e\rangle, \langle e\rangle\rangle} x^{\langle e\rangle}$

By putting a monadic second-order variable in the empty argument place and binding it with an existential quantifier, we obtain

$(\text{MORE}_2) \exists y^{\langle e\rangle} \forall_o x^{\langle e\rangle}\ y^{\langle e\rangle} \neq^{\langle\langle e\rangle, \langle e\rangle\rangle} x^{\langle e\rangle}$

It seems to me that it would be very implausible to discount this claim as irrelevant to the spirit of Absolutism on the grounds that it does not operate with the notion of *distinctness* as it occurs in (GA_1) and (MORE_1). It seems much more plausible to interpret this sentence as saying, in a sense that is relevant to the spirit of Absolutism, that the quantifier '$\forall_o x^{\langle e\rangle}$' expresses only a restricted form of (monadic, second-order) generality. If so, then (DIST) should be rejected.

 © Krämer 2016

Analogues of the ordinary identity predicate allow for the formulation of analogues of (MORE$_1$) that contradict the spirit of Absolutism.[23]

We may thus turn to the second step in my argument. I shall argue that once we are allowed to use cumulative resources, we can formulate another variant on ($\neq_1$), employing a cumulative higher-order counterpart of the identity predicate. We can then use that expression to formulate a cumulative higher-order version of (MORE$_1$) as our regimentation of (MORE). Just as we took (MORE$_2$) to contradict the spirit of Absolutism, it seems to me, we should take that regimentation to do so as well. And since it turns out to be derivable in the semantics of the cumulative higher-orderist, the introduction of cumulative resources into higher-orderism is in this way seen to undermine the spirit of Absolutism.

We form the cumulative second-order language $\mathcal{L}_{2\equiv}$ by extending $\mathcal{L}_2$ with an identity-like symbol of type $\langle[e, \langle e* \rangle], [e, \langle e* \rangle]\rangle$, which we shall write as '$\equiv$' to make it easier to distinguish from the previous ones. In contrast to these, '$\equiv$' is *cumulative* in both its argument places: it accepts both singular terms and predicates of any adicity as arguments.

What should we take '$\equiv$' to mean? For present purposes, we may either take it as a new primitive expression, subject to certain inference rules, or we may take it as being given an explicit definition. If we introduce '$\equiv$' as a primitive, it should be governed by whatever rules render it as identity-like as possible, consistent with its non-standard syntax. It seems clear, then, that it should satisfy at least the following two rules. Where 'a' and 'b' belong to $[e, \langle e* \rangle]$:

$(\equiv_1) \vdash a \equiv a$
$(\equiv_2) \vdash a \equiv b \rightarrow (\Phi \rightarrow \Phi^{a/b})$

(When 'a' and 'b' belong to different types, substituting one for the other sometimes produces ill-formed results. The rule ($\equiv_2$) is therefore to be understood as restricted to well-formed instances.[24])

[23] Could a higher-orderist dig their heels in and simply insist that (MORE$_2$) and its ilk are irrelevant to Absolutism? Although such a view would seem very unattractive, I know of no reason to think that it would have to somehow turn out to be internally incoherent. However, I think that Williamson could not happily resort to such a position, since he himself exploits the analogy between the identity predicate and its higher-order counterparts in a way similar to how I have just used it (e.g. Williamson 2013, pp. 263ff.).

[24] Absent contexts other than '$\equiv$' in which expressions of different types can be exchanged without loss of grammaticality, our rules leave open the truth-value of any specific sentences in which '$\equiv$' connects expressions of different types. We consider these in more detail in the next section.

 © Krämer 2016

If we wish instead to introduce '≡' by definition, we can extend $\mathcal{L}_{2\equiv}$ by variables of type $\langle[e, \langle e*\rangle]\rangle$ and allow these to be bound by quantifiers. We may then let '$a \equiv b$' simply abbreviate the indiscernibility condition, $\forall z^{\langle[e, \langle e*\rangle]\rangle}(z^{\langle[e, \langle e*\rangle]\rangle}(a) \leftrightarrow z^{\langle[e, \langle e*\rangle]\rangle}(b))$.[25] Given a suitable comprehension scheme for the new quantifiers, the rules $(\equiv_1)$ and $(\equiv_2)$ are then derivable.[26]

We now use '≡' to formulate a regimentation of (MORE). Let us include in $\mathcal{L}_{2\equiv}$ copies '$\forall_o x^e$' and '$\exists_o x^e$' of the first-order quantifiers of $\mathcal{L}_1$. We can then formulate a cumulative variant on $(\neq_1)$ to express a condition of distinctness from anything in the range of $\mathcal{L}_1$'s quantifier:

$(\neq_C)$ $\forall_o x^e \ldots \not\equiv x^e$

Since '≡' accepts expressions of type $\langle e\rangle$ as arguments, we can now put a monadic second-order variable in its empty argument place in $(\neq_C)$, and bind it with an existential quantifier. We then obtain the following regimentation of (MORE):

(MORE_C) $\exists y^{\langle e\rangle}\, \forall_o x^e\ y^{\langle e\rangle} \not\equiv x^e$

Given $(\equiv_1)$ and $(\equiv_2)$, this sentence is derivable by broadly standard Russell-style reasoning in second-order logic.[27] I maintain that, like (MORE_2), it is similar enough to (MORE_1) to be regarded as contradicting the spirit of Absolutism.

[25] For a related but somewhat more complicated definition of a much more general version of '≡', see Linnebo and Rayo (2012, §5) and Degen and Johannsen (2000, pp. 149f.).

[26] Absent a meaningful notion of co-extensiveness defined for $[e, \langle e*\rangle]$ and thus names and predicates alike, unlike its non-cumulative cousins, '≡' cannot be introduced as an abbreviation of a co-extensiveness condition.

[27] *Proof*: We define a first-order predicate '$r^{\langle e\rangle}$' as follows, with 'x^e' ranging over everything$_o$: $r^{\langle e\rangle}(x^e) \leftrightarrow_{\text{df}} \forall x^{\langle e\rangle}(x^{\langle e\rangle} \equiv x^e \rightarrow \neg x^{\langle e\rangle}(x^e))$. Assume for *reductio* that for some$_o$ object r^e, $r^{\langle e\rangle} \equiv r^e$. Now $r^{\langle e\rangle}(r^e) \vee \neg r^{\langle e\rangle}(r^e)$. Assume $r^{\langle e\rangle}(r^e)$. Then by definition, $\forall x^{\langle e\rangle}(x^{\langle e\rangle} \equiv r^e \rightarrow \neg x^{\langle e\rangle}(r^e))$, and so in particular, $r^{\langle e\rangle} \equiv r^e \rightarrow \neg r^{\langle e\rangle}(r^e)$. By assumption, $r^{\langle e\rangle} \equiv r^e$, so $\neg r^{\langle e\rangle}(r^e)$, contradicting our assumption of $r^{\langle e\rangle}(r^e)$. So $\neg r^{\langle e\rangle}(r^e)$. Then by definition, $\neg\forall x^{\langle e\rangle}(x^{\langle e\rangle} \equiv r^e \rightarrow \neg x^{\langle e\rangle}(r^e))$. However, assume $x^{\langle e\rangle} \equiv r^e$. Since $r^{\langle e\rangle} \equiv r^e$, it follows by $(\equiv_1)$ and $(\equiv_2)$ that $x^{\langle e\rangle} \equiv r^{\langle e\rangle}$. Since $\neg r^{\langle e\rangle}(r^e)$, by another application of $(\equiv_2)$, $\neg x^{\langle e\rangle}(r^e)$. So $x^{\langle e\rangle} \equiv r^e \rightarrow \neg x^{\langle e\rangle}(r^e)$. Since $x^{\langle e\rangle}$ was arbitrary, $\forall x^{\langle e\rangle}(x^{\langle e\rangle} \equiv r^e \rightarrow \neg x^{\langle e\rangle}(r^e))$. But then by definition, $r^{\langle e\rangle}(r^e)$. Contradiction. So $r^{\langle e\rangle} \not\equiv r^e$, and since r^e was arbitrary, $\forall_o x^e\, r^{\langle e\rangle} \not\equiv x^e$. (MORE_C) follows by existential generalization on $r^{\langle e\rangle}$.

The proof assumes that the comprehension scheme for the second-order quantifiers of $\mathcal{L}_{2\equiv}$ allows impredicative instances, including our new '≡'. Specifically, the legitimacy of the definition of '$r^{\langle e\rangle}$' and/or the subsequent existential generalization on it depend in effect on this impredicative instance of the comprehension schema: $\exists y^{\langle e\rangle}\forall x^e(y^{\langle e\rangle}(x^e) \leftrightarrow \forall z^{\langle e\rangle}(z^{\langle e\rangle} \equiv x^e \rightarrow \neg z^{\langle e\rangle}(x^e)))$. Could the higher-orderist reject such impredicative instances of comprehension? I think not; such a move would appear to undermine the whole motivation for higher-orderism, since the initial Russell-style argument depends on an impredicative instance of (CPI Formal 1).

 © Krämer 2016

5. Cumulative identity

The extent to which cumulative higher-orderism's commitment to (MORE_C) is a departure from the spirit of Absolutism depends on the strength of the analogy between '$\equiv$' and '$=$', and thus (MORE_C) and (MORE_1). Given the unfamiliarity of cumulative devices in general and '$\equiv$' in particular, it may not be very easy to get a clear sense of the strength of this analogy. This section therefore examines the behaviour of '$\equiv$' and its relation to '$=$' in more detail. I argue that there is no dissimilarity between '$\equiv$' and '$=$' that could undermine the analogy between (MORE_C) and (MORE_1).

We note first that '$\equiv$' shares the distinctive structural features of the identity predicate. In particular, it expresses an *equivalence relation* in the sense that in addition to the reflexivity rule ($\equiv_1$), symmetry and transitivity rules are derivable for 'a', 'b', 'c' in $[e, \langle e* \rangle]$:

$(\equiv_3) \vdash a \equiv b \rightarrow b \equiv a$

$(\equiv_4) \vdash (a \equiv b \ \& \ b \equiv c) \rightarrow a \equiv c$

The proofs are exactly analogous to the corresponding proofs for '$=$'.

Admittedly, the reflexivity, transitivity and symmetry of the ordinary identity relation can also be expressed in a non-schematic way by means of quantifications like $\forall x\, x = x$ and $\forall x \forall y (x = y \rightarrow y = x)$. The structural features of '$\equiv$' are not thus expressible in $\mathcal{L}_{2\equiv}$. The reason is that we do not have any variables that range over an entire cumulative type. So although we can express the reflexivity of $\equiv$ with respect to type e by '$\forall x^e\, x^e \equiv x^e$', and with respect to type $\langle e \rangle$ by '$\forall x^{\langle e \rangle}\, x^{\langle e \rangle} \equiv x^{\langle e \rangle}$', and so on, we cannot express by a single sentence the reflexivity of $\equiv$ *tout court*.

We can remove even this disanalogy by moving to a new language $\mathcal{L}_{2C}$ extending $\mathcal{L}_{2\equiv}$ by what I shall call *semantically cumulative* variables. These are variables that belong to a cumulative type $[t_1, t_2, \ldots]$ without belonging to any of the accumulated types $t_1, t_2, \ldots$ Roughly speaking, they are intended to range over the entirety of values of the variables from the accumulated types. We may use underlined lowercase letters from the end of the alphabet for these variables, and express, for example, the reflexivity of $\equiv$ in $\mathcal{L}_{2C}$ as follows:[28]

[28] Although the introduction of such variables could perhaps in principle be rejected by a cumulative higher-orderist, it is important to see how natural their introduction is once syntactically cumulative expressions have been introduced. For, absent semantically cumulative variables, we have cumulative predicates with application conditions defined for a range of items that cannot be swept out by a single variable. It seems more than odd to think that it should be impossible to add variables that can take values from the entire application range of

 © Krämer 2016

$(\equiv_5)$ $\forall \underline{x}^{[e, \langle e* \rangle]}\ \underline{x}^{[e, \langle e* \rangle]} \equiv \underline{x}^{[e, \langle e* \rangle]}$

With respect to their structural features, then, '≡' and '=' seem exactly analogous.

Moreover, whenever '≡' connects two singular terms, the resulting sentence is true iff the corresponding sentence with '=' is true. More generally, any given sentence in which '≡' connects two expressions of the same non-cumulative type is true just in case the result of replacing '≡' with a suitable non-cumulative identity-like predicate is true. That is, for τ among e, $\langle e* \rangle$:

$(\equiv_6)$ $\vdash a^\tau \equiv b^\tau \leftrightarrow a^\tau =^{\langle \tau, \tau \rangle} b^\tau$

Let us then turn to the somewhat stranger contexts of '≡' in which it connects expressions of different types; call them *cross-type identifications*. Our derivation of (MORE_C) has shown that in conjunction with the rest of the higher-orderist's logic, $(\equiv_1)$ and $(\equiv_2)$ already have substantive *general* implications concerning cross-type identifications: roughly speaking, at least one item in type $\langle e \rangle$ cannot be identified with any item in type e. The truth-values of all *specific* cross-type identifications, however, are left open by our theory. That is, where τ and σ are different types among e, $\langle e* \rangle$, no sentence of the form $a^\tau \equiv b^\sigma$ is either derivable or refutable from our rules.[29] This is exactly parallel to the situation for the logic of '=': where a and b are distinct singular terms, $a = b$ is neither derivable nor refutable from the logical rules alone. Of course, if we are given some sentences $\Phi(a)$, $\neg\Phi(b)$ as premisses, we can infer $a \neq b$ from them. But in just the same way, for a and b in $[e, \langle e* \rangle]$, and given sentences $\Phi(a)$, $\neg\Phi(b)$ as premisses, we may also infer $a \not\equiv b$ from them. So far, then, no relevant disanalogy between '=' and '≡' has emerged.

---

the predicates. I shall henceforth assume that semantically cumulative variables are no more problematic than syntactically cumulative expressions.

[29] It is clear that our rules do not permit the derivation of any theorem of the form $a \equiv b$ where a and b are distinct expressions. *A fortiori*, they do not permit the derivation of any such theorem where a and b belong to different types. It may not be as obvious that $(\equiv_2)$ does not, as it stands, allow the derivation of any negation of a cross-type identification. But note that so far, '≡' is the *only* cumulative predicate in our language. So any well-formed cross-type instance of $(\equiv_2)$ will be a formula like '$a \equiv b \rightarrow (b \equiv c \rightarrow a \equiv \varsigma)$', where the substitution in the consequent occurs in the scope of '≡'. As a result, we could only obtain the negation of a cross-type identity from such a premiss given another negation of a cross-type identity to start with. For essentially the same reason, we also cannot derive negations of cross-type identifications appealing to the proposed indiscernibility definition of '≡'.

 © Krämer 2016

Perhaps one might be tempted to argue for such a disanalogy along the following lines.[30] A cumulative counterpart of the identity predicate like '≡' is most naturally thought of on the model of the *disjunction* of the relevant non-cumulative identity-like predicates. Now, when we think of '≡' in this way, then it *trivially* produces a false sentence whenever it is fed expressions of different types as arguments. But that makes the crucial claims of distinctness from which (MORE_C) is inferred importantly dissimilar to the claims of distinctness from which (MORE_1) is inferred. For the latter claims of distinctness are not trivial in this way. And since (MORE_C) is now seen to be merely an immediate consequence of a perfectly trivial claim, it would be implausible to consider it as contradicting the spirit of Absolutism. For surely it would be disingenuous to interpret the spirit of Absolutism in such a way that it is straightforwardly inconsistent with a mere triviality.

My response is that the envisaged disjunctive understanding of '≡' is incompatible with how I have introduced the expression, and that an understanding that is in line with how I have introduced '≡' does not trivialize cross-type identifications in the way described. I suggested two alternative ways to introduce '≡'. The first is to take it as primitive, subject to the inference rules ($\equiv_1$) and ($\equiv_2$). These rules do not trivialize cross-type identifications, but tie them to predications involving the relevant terms. As a result, there are non-trivial considerations that can be brought to bear on the question of the truth or falsity of a given cross-type identification.

Consider first the question of the identity between objects a and b. In order to decide the question, we may ask whether there is a predicate that, under a uniform interpretation, is defined for both a and b, and applies to a but not to b. Since the cumulative higher-orderist allows cumulative predicates, a counterpart of that consideration applies to questions of cross-type identity. For instance, to decide whether $a^e \equiv b^{\langle e \rangle}$, we may ask whether there is a predicate of type $\langle [e, \langle e \rangle] \rangle$—whose application conditions are accordingly defined for both a^e and $b^{\langle e \rangle}$—which, under a uniform interpretation, applies to a^e but not to $b^{\langle e \rangle}$.

The second way of introducing '≡' I offered is through an explicit definition in terms of the corresponding indiscernibility condition. This definition does not trivialize cross-type identifications either, but ties them to the existence of discriminating properties. To

[30] Thanks here to Nick Haverkamp.

 © Krämer 2016

decide whether a given cross-type identification is true, we may ask whether there is a property of type $\langle [e, \langle e \rangle] \rangle$—whose exemplification conditions are accordingly defined for both a^e and $b^{\langle e \rangle}$—which a^e has but $b^{\langle e \rangle}$ does not. In the same way, the existence of discriminating properties bears on questions of first-order identity: we ask whether there is a property whose exemplification conditions are defined for both a and b and which a has and b does not.

Finally, it is not obvious whether some cross-type identifications might not be true.[31] Consider some way for a thing to be, say, wise, and the corresponding property, wisdom, conceived of as a bona fide object, that is, something properly designated in a formal language by a singular term. It is not absurd to think that if '$\equiv$' is to be the closest thing to the identity predicate that is defined for a range comprising both objects and ways for objects to be, then it should be the case that wisdom $\equiv$ is wise.[32] Such a view might allow a cumulative higher-orderist to hold that most of the ways for things to be recognized in $\mathcal{L}_{2\equiv}$ are 'identical' to things in the range of $\mathcal{L}_1$'s quantifier—the exception being the Russell-style ways for things to be. Indeed, a non-classical cumulative higher-orderist who endorses a naive comprehension schema for properties at the cost of some truths of classical logic could in this way argue for the negation of (MORE_C):

(GA_C) $\forall \underline{y}^{[e, \langle e* \rangle]} \exists_o x \; \underline{y}^{[e, \langle e* \rangle]} \equiv x$

Whatever the overall merits or demerits of such a view, it would thereby underwrite a stronger form of Absolutism than is consistent with a classical cumulative higher-orderism such as Williamson's.

I conclude that '$\equiv$' is not problematically dissimilar to '$=$', and that the analogy between (MORE_1) and (MORE_C) is accordingly strong enough that (MORE_C) should be taken to contradict the spirit of Absolutism.

[31] It should be noted that the option of holding things of different types in general distinct is plausible at most for what we may call *pure* types, that is, types that are neither cumulative nor derived from cumulative types. If we allow semantically cumulative variables, we of course obtain some examples of true cross-type identifications. In particular, we should have the result that, for example, $\forall x^e \exists \underline{y}^{[e, \langle e* \rangle]} \; x^e \equiv \underline{y}^{[e, \langle e* \rangle]}$. Moreover, for higher-level analogues of '$\equiv$' which also apply to expressions of functional types derived from cumulative types, we shall also have true cross-type identifications where neither argument is of a cumulative type. For example, for the counterpart of '$\equiv$' in type $\langle [\langle e \rangle, \langle [e, \langle e \rangle] \rangle], [\langle e \rangle, \langle [e, \langle e \rangle] \rangle] \rangle$ we should have that $\forall x^{\langle e \rangle} \exists y^{\langle [e, \langle e \rangle] \rangle} \; x^{\langle e \rangle} \equiv y^{\langle [e, \langle e \rangle] \rangle}$.

[32] The details of how nominalizations behave in natural language may lend further support to this idea; see Liebesman (2015, esp. §3).

 © Krämer 2016

6. Expressive difficulties

One of Williamson's arguments against Relativism is related to the peculiarity we noted in §2 concerning the most natural ways an Absolutist might attempt to express their view. Prima facie, the obvious way to do that is to utter a sentence like 'It is possible to quantify over absolutely everything'. Now if Absolutism is true, then we can interpret the Absolutist's use of the quantifier phrase 'absolutely everything' in the absolutist way they intend, and under such an interpretation the Absolutist has said something relevant and true. But if Absolutism is false, then we cannot so interpret the utterance, and our Absolutist, far from making a relevant claim that is unfortunately not true, has made a true but irrelevant claim. While that is perhaps a somewhat strange dialectical situation, it seems that at least by their own lights, the Absolutist can express their intended thesis.

Williamson (2003, §5) suggests that the Relativist is in a considerably worse position. The idea is this. Suppose the Relativist attempts to state his view by uttering the negation of the Absolutist's sentence, that is, 'It is impossible to quantify over absolutely everything'. That sentence can express what the Relativist intends it to express only on the assumption that Relativism is false. And in that case, what the sentence expresses is also false, as one would have hoped. But if Relativism is true, the sentence simply says something unintended. Whatever restricted domain the quantifier 'absolutely everything' is interpreted as ranging over, the sentence then says that quantification restricted in that way is impossible. But that is not something the Relativist wants to proclaim to be impossible, and rightly so, for it evidently is not impossible. It seems that if the Absolutists cannot help but say something true, but at least say what they want if their view is right, the Relativists cannot help but say something false, and say what they want only if their view is wrong.

Of course, even if successful, this argument does not show that standard Relativism is incoherent. It shows only that a natural first idea for stating the view is incoherent. Relativists might respond in two ways. They might simply resist the urge to produce a general claim supposed to capture their view, limiting themselves to claims like

$$(\text{MORE}_1)\ \exists y \forall_o x\ x \neq y$$

describing individual languages like $\mathcal{L}_1$ as expressing only a restricted form of generality. Although these claims seem to instantiate a common pattern, crying out for generalization, Relativists might

 © Krämer 2016

simply reject any suggestion that there is a true generalization of which all these claims are mere special cases.[33] Admittedly, while internally coherent, this kind of quietist position may be less than fully satisfactory. Alternatively, Relativists can try to devise some other way of formulating a general thesis that can capture their view without collapsing into incoherence.[34]

Whether or not Relativists can find a convincing solution to the problem, it appears that cumulative higher-orderists face exactly the same sort of difficulty. Their proposal requires that from any legitimate language, we can move to a metalanguage that includes quantifiers of an order higher than any order of quantification found in the object language. So like Relativists, they seem committed to a kind of inexhausibility thesis; only the higher-orderist's thesis concerns the entire hierarchy of orders of quantification. The most natural way to attempt to express the view is by uttering a sentence like 'It is impossible to quantify over absolutely everything in the entire hierarchy of higher and higher orders of quantification'. Clearly, this sentence is no better off than the Relativist's doomed 'It is impossible to quantify over absolutely everything'.

Of course, this does not show that cumulative higher-orderism is incoherent. It shows only that a natural first idea for stating the relativist component of the view is incoherent. Like Relativists, higher-orderists might respond in two ways. They might simply resist the urge to produce a general claim supposed to capture the relativist element of their view, limiting themselves to claims like

$$(\mathrm{MORE}_C)\ \exists y^{\langle e\rangle} \forall_o x^e\ y^{\langle e\rangle} \not\equiv x^e$$

describing individual languages like $\mathcal{L}_1$ as expressing only a limited form of generality. Although these claims seem to instantiate a common pattern, crying out for generalization, higher-orderists might simply reject any suggestion that there is a true generalization of which all these claims are mere special cases. Again, the view seems to be internally coherent, but to the same extent as its Relativist cousin, it also seems less than fully satisfactory. Alternatively, higher-orderists can try to devise some other way of formulating a general thesis capturing their intended form of relativism. It is not obvious how they might do so, but if they manage it, that would answer the objection just presented. However, it is to be expected

[33] Button (2010) puts forth a view like this.

[34] See Fine (2006) for one proposal.

 © Krämer 2016

(though I cannot prove) that if any such means were to be found, it would also give the standard Relativist a way of formulating their thesis without falling into incoherence.

I conclude that, as far as adequate formulation of their overall view is concerned, standard Relativists and cumulative higher-orderists are in a very similar situation.[35]

7. Interpretative limitations

Williamson (2003, pp. 415ff., 435) also criticizes Relativism for painting an unattractive picture of parts of logical, metaphysical and mathematical discourse. As I mentioned in the introduction, such discourse appears to provide us with numerous examples of utterances intended as absolutely general. As a result, the interpretations of such utterances that the Relativist can offer seem, from a pre-theoretic standpoint, quite weird and implausible. In this section, I investigate whether the interpretations offered by the cumulative higher-orderist should be considered similarly weird and implausible.

Let me first try to bring out as clearly as I can why the Relativist's interpretations seem weird.[36] Consider any of the following sentences:

(6) Everything is self-identical.
(7) If something x is identical to something y, then this is necessarily so.
(8) Everything is necessarily identical to something.
(9) Everything is either abstract or concrete.

[35] The present criticism of *cumulative* higher-orderism is somewhat similar to criticisms advanced in Linnebo (2006, §4) against *non-cumulative* higher-orderism. Linnebo argues that the higher-orderist is committed to claims such as that there are infinitely many different kinds of semantic value, but cannot by his own lights give proper expression to these claims. He maintains that this limitation is similar to that afflicting Generality Relativism. In an earlier paper (Krämer 2014), I have offered a response to Linnebo's arguments on behalf of the non-cumulative higher-orderist; unfortunately, it does not carry over to protect cumulative higher-orderism against the above argument.

[36] Although I myself agree that they do seem weird, I have encountered some resistance to this view in discussion. So it may be worth stressing two points. Firstly, it is important to guard against a misunderstanding. As I see it, the issue concerns the plausibility of certain interpretations from a *pre-theoretic* standpoint. Among other things, this means that 'interpretation' here must not be understood in any theoretically loaded sense like *assignment of semantic values*, or some such, but is to be taken in an intuitive, pre-theoretic sense. Secondly, for the purposes of my argument, it does not matter much whether the following considerations do show that the Relativist's interpretations should seem implausible. For my claim will simply be that if they do, then similar considerations establish an analogous, if less drastic, verdict on the cumulative higher-orderist's interpretations.

 © Krämer 2016

From the point of view of a Relativist's metalanguage and semantics, utterances of any of (6)–(9) can be interpreted only as restricted to some less than all-inclusive domain *D*. In typical cases, any such interpretation seems to simply be a *mis*interpretation. Consider the formalization

(10) $\forall x\,(Ax \vee Cx)$

of (9) in a formal first-order language $\mathcal{L}_1$. Assume we have formulated a Relativist semantics in a first-order metalanguage $\mathcal{L}_2$, which we may suppose includes the predicates 'is abstract' and 'is concrete' with their usual meanings. Now consider the following two sentences of $\mathcal{L}_2$:

(11) $\forall x$ (x is concrete $\vee$ x is abstract)

(12) $\forall x \in D$ (x is concrete$_D$ $\vee$ x is abstract$_D$)

where D is the domain with respect to which we are interpreting $\mathcal{L}_1$, and 'abstract$_D$' and 'concrete$_D$' mean the same as their subscript-free counterparts except that their application conditions are defined only over D. Evidently, the generalization expressed by (12), as well as every generalization expressed by a version of (12) in which some other set is referred to in place of D, bears an interesting relationship to the generalization expressed by (11). Roughly speaking, (12) is obtained from (11) by replacing the quantifier '$\forall x$' with a proper restriction of it, and replacing the predicates with counterparts having accordingly restricted application conditions. I shall say the generalization expressed by (12) is a *mere restriction* of that expressed by (11), and conversely that the latter is a *mere expansion* of the former.

The important point is that the claim expressed by (12), because it is a mere restriction of that expressed by (11), seems a strange target for metaphysical inquiry, and relatedly, a strange claim to put forth as the upshot of a metaphysical investigation. The claim expressed by (11) seems a much more interesting and natural claim to focus on. At least at a first, and again, perhaps somewhat naive glance, it therefore seems implausible to interpret a metaphysician using the $\mathcal{L}_1$-sentence (10) to put forth (part of) his metaphysical theory as having endorsed the claim expressed by (12), not (11). Of course, if we try to specify, in a metalanguage for our metalanguage, an intended interpretation of (11), it too will turn out to express a mere restriction of a yet more encompassing counterpart. In this way, saying that everything is abstract or concrete becomes something of a metaphysisyphean task.

To illustrate the point another way, imagine a necessitist like Williamson uttering (8). In effect, the Relativist would interpret him as saying that everything *except perhaps some things outside D* is

 © Krämer 2016

necessarily (identical to) something. This seems very strange. From a naive point of view, one expects the Relativist to then ask our necessitist what his view is with respect to the things outside *D*. The kind of conversation that would then unfold seems not worth having. N: 'I think that *everything*, in or outside *D*, is necessarily something.' R: 'Oh, I see, you think that everything in your domain of discourse *D*′ is necessarily something. But what about the things outside *D*′?' N: 'Yes, them too. Indeed, *everything*, and so in particular everything outside *D*′, is necessarily something.' R: 'I see, so everything in *D*″ …'

Thankfully, disputes in logic and metaphysics rarely take this shape. But it is not obvious why interpreting speakers in the Relativist's way should not invite this kind of tiring sequence of questions and responses. Relatedly, it is not obvious why the Relativist's interpretations of the pertinent utterances are not cases of strange misinterpretation. So at first glance, it seems that Relativism paints a somewhat disconcerting picture of significant parts of logico-metaphysical inquiry.

However, it seems to me that the same kind of problem also arises for the cumulative higher-orderist, although only in a somewhat less dramatic form. Specifically, for some universal generalizations of $\mathcal{L}_1$, it seems that, from the perspective of the cumulative metalanguage $\mathcal{L}_{2C}$, they express mere restrictions of more encompassing claims. This is so in particular for $\mathcal{L}_1$-generalizations involving only logical vocabulary.

Consider the $\mathcal{L}_1$-sentence

(13) $\forall x\ x = x$

On the cumulative higher-orderist's semantics given in $\mathcal{L}_{2C}$, it expresses the claim that

(14) $\forall x^e\ x^e = x^e$

Now, as we have seen, using cumulative resources, we can express a version of the notion of identity that is defined both for objects and for ways for objects to be. That notion of $\equiv$ seems to relate to the ordinary notion of identity in much the same way as the notion of abstractness relates to that expressed by 'abstract$_D$' above. The ordinary identity predicate '$=$' means the same as '$\equiv$', except that its application conditions are defined only for objects.

Moreover, the semantically cumulative quantifier '$\forall \underline{x}^{[e,\ \langle e* \rangle]}$' of $\mathcal{L}_{2C}$ seems to relate to '$\forall x^e$' as expansion to restriction in much the same

 © Krämer 2016

way as '$\forall x$' above does to '$\forall x \in D$'. Indeed, from the perspective of $\mathcal{L}_{2C}$, we could also say that (13) expresses the claim that[37]

(15) $\forall \underline{x}\ (\mathrm{OBJ}(\underline{x}) \rightarrow \underline{x} \equiv_{\mathrm{OBJ}} \underline{x})$

where '$\mathrm{OBJ}(\underline{x})$' is defined by '$\exists y^e\ \underline{x} \equiv y$', and thus applies to all and only objects, and '$\underline{x} \equiv_{\mathrm{OBJ}} \underline{y}$' is defined by '$\underline{x} \equiv \underline{y}$' except that it is defined only for objects.

In this way, we can formulate mere expansions for any given universal generalization in $\mathcal{L}_1$ as long as it contains only first-order predicates that have natural higher-order analogues. Clearly, this is not the case for *every* generalization. For example, it does not seem as though the abstract–concrete distinction has any obvious counterparts at higher orders of quantification. Nevertheless, it holds for quite a few interesting and contentious logico-metaphysical theses, such as the necessity of identity and existence. More cases show up when we consider a second-order language as object language. They include, for example, the claim that nothing has contradictory properties, which we can formulate as $\forall x^e \neg \exists x^{\langle e \rangle} (x^{\langle e \rangle}(x^e) \,\&\, \neg x^{\langle e \rangle}(x^e))$, or the conclusion of the Russell Paradox, which may be expressed as $\neg \exists z^{\langle e,\, e \rangle} \forall x^{\langle e \rangle} \exists x^e \forall y^e\ (x^{\langle e \rangle}(y^e) \leftrightarrow z^{\langle e,\, e \rangle}(x^e, y^e))$.[38]

It would be an exaggeration to claim in conclusion that with respect to the difficulty of interpreting apparently absolutely general discourse in logic, mathematics or metaphysics, a cumulative higher-orderist is in as worrying a position as the Relativist. The cases in which they may be thought to interpret speakers in an implausibly restricted fashion are less widespread than those arising for Relativism, and it is not obvious that the restricted interpretations are as implausible-looking as those given by the Relativist. Nevertheless, the difference appears to be one of degree, and a smaller one than one might have expected.[39]

[37] For readability, I have here dropped the type-superscript $[e, \langle e* \rangle]$ from the underlined, cumulative variables.

[38] A detailed assessment and comparison of the seriousness of the problem for Relativism and cumulative higher-orderism is beyond the scope of this paper, but the following consideration may indicate a reasonable way to start. Suppose someone makes an utterance about a topic *t*, and consider the distinctive 'domain-expanding' moves then canvassed by Relativism and cumulative higher-orderism, respectively. We may then ask whether the additional expressive resources available after the move enable us to exclude ways for the world to be *with respect to t* that we could not previously exclude. If it turns out that the range of cases in which this is so is significantly smaller for cumulative higher-orderism than it is for Relativism, this would seem to speak in favour of the former view. (Thanks to an anonymous referee for suggesting this way of viewing the matter.)

[39] There may be further examples of problems arising for higher-orderism which seem akin to problems of Generality Relativism. In particular, Linnebo and Rayo (2012, §7) argue that

 © Krämer 2016

8. Conclusion

The attempt to develop a generalized semantic theory in a first-order language for an object language that expresses absolute generality runs into a version of Russell's Paradox. Standard Relativism concludes that absolute generality is impossible. Williamson (2003) criticizes that view on a number of grounds, such as that it is by its own lights not properly expressible, and that it yields an unsatisfactory picture of parts of logical, metaphysical and mathematical inquiry. To avoid Relativism, Williamson appeals to a hierarchy of higher and higher orders of quantification, which, in his recent book, he extends to transfinite and cumulative orders of quantification. I have argued that the emerging view has implications strongly reminiscent of the distinctive commitments of standard Relativism, and thus gives up on much of the spirit of Absolutism. Moreover, versions of the two criticisms of Relativism just mentioned also apply on the cumulative higher-order view.

Cumulative higher-orderism is not thereby shown to be wrong, of course. (Indeed, I incline to think it is correct.) What I hope to have shown, however, is that it does not let us be as much of a Generality Absolutist as we might have thought and hoped. Even according to cumulative higher-orderism, there is not just everything. There is everything, and then some.[40]

---

higher-orderists are committed to a strictly open-ended type-theoretic hierarchy, and that this appears to face problems similar to the Relativist's open-ended ontological hierarchy. It may be worth stressing that the claims and arguments of the preceding two sections are independent of the issue of open-endedness. They turn crucially on the *cumulative* nature of the type-theoretic hierarchy rather than its size. Specifically, it is thanks to cumulative types that a regimentation of the distinctive relativist claim (MORE) can be formulated and proved in cumulative higher-order logic, and that what appear to be mere expansions of lower-order generalizations become expressible at higher cumulative orders. (Thanks to an anonymous referee for pressing me on this.)

[40] Earlier versions of this paper were presented at the Hamburg Summer School 2012, at which Timothy Williamson presented his book, *Modal Logic as Metaphysics*, and at a workshop on Typed and Untyped Approaches in Semantics in Oslo. I am grateful to the audiences at both occasions for helpful discussion, and especially to Timothy Williamson for extremely helpful and encouraging comments. Many thanks are also due to Nick Haverkamp, Stefan Roski, Benjamin Schnieder, Richard Woodward, and all fellow members of the Hamburg *phlox* research group, as well as two referees and the editor of *Mind* for helpful comments and criticism. The research for this paper was carried out while I was employed in the DFG/ANR funded project Nominalizations, and I gratefully acknowledge that support.

 © Krämer 2016

References

Boolos, George 1984: 'To Be is to be a Value of a Variable (or to be Some Values of Some Variables)'. *Journal of Philosophy*, 81, pp. 430–49.

——1985: 'Nominalist Platonism'. *Philosophical Review*, 94, pp. 327–44.

Button, Tim 2010: 'Dadaism: Restrictivism as Militant Quietism'. *Proceedings of the Aristotelian Society*, 110, pp. 387–98.

Degen, Wolfgang, and Jan Johannsen 2000: 'Cumulative Higher-Order Logic as a Foundation for Set Theory'. *Mathematical Logic Quarterly*, 46, pp. 147–70.

Fine, Kit 2006: 'Relatively Unrestricted Quantification'. In Rayo and Uzquiano 2006b, pp. 20–44.

Florio, Salvatore 2014: 'Unrestricted Quantification'. *Philosophy Compass*, 9(7), pp. 441–54.

Krämer, Stephan 2014: 'Semantic Values in Higher-Order Semantics'. *Philosophical Studies*, 168, pp. 709–24.

Liebesman, David 2015: 'Predication as Ascription'. *Mind*, 124, pp. 517–69.

Linnebo, Øystein 2006: 'Sets, Properties, and Unrestricted Quantification'. In Rayo and Uzquiano 2006b, pp. 149–78.

Linnebo, Øystein and Agustín Rayo 2012: 'Hierarchies Ontological and Ideological'. *Mind*, 121, pp. 269–308.

Prior A. N. 1971: *Objects of Thought*. Edited by P. T. Geach and A. J. P. Kenny. Oxford: Clarendon Press.

Rayo, Agustín 2006: 'Beyond Plurals'. In Rayo and Uzquiano 2006b, pp. 220–54.

Rayo, Agustín and Gabriel Uzquiano 1999: 'Toward a Theory of Second-Order Consequence'. *Notre Dame Journal of Formal Logic*, 40(3), pp. 1–11.

——2006b: 'Introduction'. In Rayo and Uzquiano 2006b, pp. 1–20.

——2006a: *Absolute Generality*. Oxford: Oxford University Press.

Rayo, Agustín, and Stephen Yablo 2001: 'Nominalism through De-Nominalization'. *Noûs*, 35(1), pp. 74–92.

Williamson, Timothy 2003: 'Everything'. *Philosophical Perspectives*, 17, pp. 415–65.

——2013: *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

© Krämer 2016

Naïve Realism In Kantian Phrase

Anil Gomes
Trinity College, University of Oxford
anil.gomes@trinity.ox.ac.uk

Early twentieth-century philosophers of perception presented their naïve realist views of perceptual experience in anti-Kantian terms. For they took naïve realism about perceptual experience to be incompatible with Kant's claims about the way the understanding is necessarily involved in perceptual consciousness. This essay seeks to situate a naïve realist account of visual experience within a recognisably Kantian framework by arguing that a naïve realist account of visual experience is compatible with the claim that the understanding is necessarily involved in the perceptual experience of those rational beings with discursive intellects. The resultant view is middle way between recent conceptualist and non-conceptualist interpretations of Kant, holding that the understanding is necessarily involved in the kind of perceptual consciousness that we, as rational beings, enjoy whilst allowing that the relations of apprehension which constitute perceptual consciousness are independent of acts of the understanding.

1. Introduction

Our textbook nativity stories often trace the foundation of analytic philosophy back to the rejection of neo-Hegelian idealism in Cambridge at the start of the twentieth century. It is less often appreciated that this rejection was also a rejection of *Kantian* idealism and, in particular, a rejection of Kant's account of the role the mind plays in shaping perceptual access to the world. The birth of analytic philosophy in both Oxford and Cambridge involved a sustained engagement with Kant's works and many of those involved in this undertaking – G.E. Moore, John Cook Wilson, H.A. Pritchard – rejected Kant's account of perceptual consciousness as incompatible with a straightforward realism about our perceptual access to empirical objects.[1]

[1] The clearest statement of this rejection is in Prichard's *Kant's Theory of Knowledge* (Prichard 1909). Both Cook Wilson and Prichard lectured on Kant's *Critique of Pure Reason* (Prichard 1909, p. iii) and Cook Wilson's writings (Cook Wilson 1926) contain a number of discussions of Kantian themes. Indeed, A.S.L. Farquaharson, in his memoir of Cook Wilson, goes so far as to trace the origins of the Oxford strand of early analytic philosophy back to questions posed by a student during Cook Wilson's lectures on Kant (Cook Wilson 1926,

Mind, Vol. 126 . 502 . April 2017 **© Gomes 2016**
doi:10.1093/mind/fzw009 Advance Access publication 24 December 2016

These Oxford and Cambridge realists instead endorsed a view on which perceptual consciousness is constituted by a primitive relation of acquaintance or apprehension, one which presents the world to the perceiver by way of non-conceptual sensory capacities and without any input or intellectual activity from the apprehending mind (Cook Wilson 1926; Prichard 1909; Moore 1903b; Russell 1910, 1912). They differed, at least after some vacillation in Cambridge, as to the second relatum of this relation. The Oxford Realists held that this relation of apprehension is one that subjects stand in to ordinary empirical objects and their properties; the Cambridge Realists – motivated by considerations involving illusions and mistaken appearances – recanted this view in favour of the claim that it is a relation that subjects stand in to mind-independent but non-physical sense-data (Moore 1913; Broad 1923; see also Martin 2003; Hatfield 2013).

Let us call the view which remained in Oxford *naïve realism*. It encompasses a tradition in the philosophy of perception which stretches from Cook Wilson and Prichard to the recent writings of Mike Martin and Charles Travis amongst others. Naïve realist accounts of perceptual experience have become central in recent debates about our perceptual access to the world (Martin 2002a; Campbell 2002b; Brewer 2011; Travis 2013a). And the claim of these early realists was that naïve realism is incompatible with Kant's views on the nature of perception.

This may sound like common sense. After all, we all know that Kant is a transcendental idealist – and whatever transcendental idealism is, it doesn't sound very naïve, nor very realist. My aim in this paper is to show that this reaction is mistaken. Kant's account of perception has much in common with naïve realism, and the early-twentieth-century naïve realist conception of visual perception can be incorporated within a recognisably Kantian framework.

Why think that naïve realism is incompatible with Kant's views on the nature of perception? §2 will characterise naïve realism in more detail. For now, the following rough characterisation will suffice: naïve realism takes visual perceptual consciousness to be constituted by

---

p. xix). The 1897 and 1898 dissertations which Moore submitted for the Trinity College Prize Fellowship examinations (Moore 2011) include sustained discussion of Kant's idealism and mark the start of Moore's conversion from idealism to realism; his 'Refutation of Idealism' (1903b) and 'Kant's Idealism' (1903a) – the title of the former containing a somewhat ironic allusion to Kant – include a rejection of Kantian claims about perception as incompatible with realism. See Hanna (2001) for further discussion of the anti-Kantian character of early analytic philosophy.

 © Gomes 2016

primitive relations of apprehension. Subjects can stand in these relations without representing the world as being a particular way. The objects of these relations are the ordinary empirical objects which populate the world around us. And in virtue of standing in these relations, perceptual consciousness presents us with the empirical world.

This account of visual perceptual consciousness has important affinities with Kant's discussion of our perceptual access to objects. I will outline the case in more detail in §3 below, but for now note that, according to the first *Critique*, there are two distinct faculties involved in human cognition: a passive faculty of sensible reception and an active faculty of the understanding (A50/B74). The role of sensibility is to provide us with objects; it does this through intuitions [*Anschauungen*] which relate us immediately to the objects of experience. In contrast, the role of the understanding is to enable us to think of objects, and it does this through the use of concepts [*Begriffe*] (A19/B33). Both capacities are drawn upon in our cognizing the world: 'without sensibility no object would be given to us, without understanding no object would be thought' (A51/B76). But 'these two faculties cannot exchange their functions. The understanding is not capable of intuiting anything, and the senses are not capable of thinking anything. Only from their unification can cognition arise' (A51/B76).

This division of our cognitive faculties appears to imply that sensibility alone is sufficient for the perceptual presentation of objects in intuition: 'objects are given to us by means of sensibility, and it alone yields us intuitions' (A19/B34). And 'intuition by no means requires the functions of thinking' (A91/B123); it is given 'prior to all thinking' (B132). This suggests that sensibility presents us with objects without any input from our conceptual faculties. Although this doesn't *entail* a naïve realist account of perceptual consciousness, it at least seems compatible with it and the potential similarities are intriguing.

So whence arises the incompatibility? The trouble arises because Kant appears to take back this initial characterisation of sensibility – or, at least, severely reinterpret it – in the Transcendental Analytic section of the first *Critique*. In contrast with sensibility, the understanding is a faculty for judging or thinking (A69/B94); it does this by means of concepts which represent what is common to more than one object (A19/B33); and all judgements are 'functions of unity among our representations' (A69/B94). But 'the same function which gives unity to the various representations in a judgment also gives unity to the mere synthesis of representations in an intuition' (A79/B105), and

 © Gomes 2016

'all synthesis, through which even perception itself becomes possible, stands under the categories' (B161). Thus perceptual consciousness itself involves acts of synthesis undertaken by the understanding and guided by conceptual capacities. And if perceptual consciousness involves acts of synthesis undertaken by the understanding, then purely sensible relations alone cannot suffice for perceptual consciousness, contrary to the tenets of naïve realism.

The Oxford Realist response to this supposed incompatibility is to reject Kant's theory of perceptual experience. An alternative response is to deny that Kant held the understanding to be constitutive of perceptual consciousness. According to recent *non-conceptualist* interpretations of Kant (Hanna 2001, 2005; Allais 2009, 2012), Kant denies that the application of concepts is necessary for the perceptual presentation of outer particulars. One striking motivation for endorsing such a view consists in the ascription to Kant of some form of relational or naïve realist account of perception. As Allais puts it, 'although he does not explicitly situate his view in terms of the theories of perception we discuss today, [Kant] clearly commits himself to a key part of the direct realist or relational position' (2011, p. 382). The reasoning follows that of the Oxford Realists: if Kant held that sensibility alone suffices for perceptual consciousness, he cannot have thought the involvement of conceptual capacities necessary for such consciousness.

These non-conceptualist interpretations make naïve realism compatible with Kant by denying that Kant held the understanding to be necessarily involved in perceptual consciousness. On this picture, the understanding is not involved in 'the basic capacity to perceive particulars but … [only in] the capacity to apply empirical concepts in judgements' (Allais 2012, p. 103). Such an account brings Kant's theory of perceptual experience very close to the naïve realism of the Oxford Realists: perceptual consciousness presents us with outer particulars without any involvement of the understanding. But it does so by denying any necessary involvement of the understanding in perceptual consciousness.[2]

My aim in this paper is to present an alternative way to reconcile naïve realism and Kant's theory of perceptual experience: one which holds, with naïve realism, that visual perceptual consciousness is

[2] Such views can allow the understanding to be causally relevant in bringing about perceptual consciousness. And they can allow occasions on which perceptual consciousness is cognitively penetrated by the understanding. But they deny that there is anything necessary about such penetration. See §3 below.

 © Gomes 2016

constituted by relations of apprehension but claims also that the understanding is necessarily involved in the perceptual experience of those rational beings with discursive intellects. The resultant view is middle way between recent conceptualist and non-conceptualist interpretations of Kant, holding that the understanding is necessarily involved in the kind of perceptual consciousness that we, as rational beings, enjoy whilst allowing that the relations of apprehension which constitute perceptual consciousness are independent of acts of the understanding.

The title of the paper comes from P.F. Strawson's 1979 paper, 'Perception and Its Objects'. In that paper, Strawson undertakes to provide a strict account of visual experience which in no way 'distort[s] or misrepresent[s] the character of that experience as we actually enjoy it' (Strawson 1979, p. 127). The resulting account of our ordinary conception of visual experience has been claimed to have affinities with both naïve realism *and* concept-involving accounts of perception.[3] And Strawson summarises it in the following words: 'mature sensible experience (in general) presents itself as, in Kantian phrase, an *immediate* consciousness of the existence of things outside us' (p. 132, emphasis in original). The words 'in Kantian phrase' are often excised in quotation. But they are my spur in what follows. For they offer the promise that a naïve realist account of visual perception can be accommodated within a Kantian account of experience, one which takes the understanding to be actively involved in perceptual consciousness contrary to both the Oxford realists and the non-conceptualist interpreters of Kant.

I'll proceed as follows. In §2 I set out a more specific characterisation of naïve realism and clarify its supposed tension with Kant's account of perceptual experience. In §3 I provide some reason for thinking both that Kant's account of sensibility can profitably be read on the model of a naïve realist account of visual perception and that Kant was committed to the involvement of the understanding in perceptual consciousness. In §4 I suggest a way in which a Kantian account of experience can make room for the naïve realist picture of visual perception by showing how naïve realism about visual experience is compatible with the essential involvement of the

[3] See Crane (2005a, pp. 245-57); Nudds (2009, p. 334 fn. 2) for references to Strawson in the context of motivating naïve realism; see Strawson (1979, p. 129) for the claim that 'sensible experience is permeated by concepts'; and cf. Strawson (1992, p. 62).

 © Gomes 2016

understanding in the perceptual consciousness of rational beings.[4] And in §5 I consider the relation of this Kantian form of naïve realism to Kant's project in the first *Critique*.

The operative term here is *Kantian*: I do not take myself to show that these are Kant's ways of thinking about the issues. Contemporary ethics distinguishes between *Kant's* ethics and *Kantian* ethics (O'Neill 1989, Part III; Wood 2008, ch. 1.1). Kant's ethics is 'the theory that Kant himself put forward' whilst Kantian ethics is 'an ethical theory formulated in the basic spirit of Kant, drawing on and acknowledging a debt to what the author of the theory takes to be his insights in moral philosophy' (Wood 2008, p. 1). The account of visual experience motivated in this paper is put forward in the spirit of a *Kantian* approach to the philosophy of perception. What I hope to provide are some suggestions for how the naïve realist insight can be accommodated within a recognisably Kantian framework for thinking about the distinction between sensibility and the understanding and the involvement of each faculty in perceptual consciousness.

One caveat before I begin: I will conduct this discussion within a framework of empirical realism and put to one side issues about how my proposal relates to transcendental idealism. This is not because I think the relation unimportant. Rather, I take it that transcendental idealism is Kant's explanation for certain features of our empirical cognitive life and thus that we can investigate Kant's account of how our intellectual capacities are active in perceptual consciousness whilst remaining neutral on the question of whether this aspect of cognition requires a transcendental idealist explanation.

2. Naïve Realism

Some definitions. The *phenomenal properties* of an experience are those in virtue of which there is something it is like to have the experience. The *phenomenal character* of an experience consists of its phenomenal properties. We type visual experiences by their phenomenal character: two visual experiences are of the same fundamental kind if and only if they have the same phenomenal character (Soteriou 2005, p. 194). *Relational* theories of visual experience are those on

[4] I confine myself here to visual perception. Kant draws distinctions between the five senses in the *Anthropology*, in particular distinguishing the objective senses of touch, sight and hearing from the subjective senses of taste and smell (*An* 7:154). These differences may be important in thinking about what a Kantian account of each sense would look like and I do not want to prejudge that issue here.

 © Gomes 2016

which the phenomenal properties of the experiences involved in vision essentially involve non-representational relations of apprehension in which subjects stand to objects. *Representational* theories of visual experience are those on which the phenomenal properties of visual experiences essentially involve representational properties. The qualifier 'non-representational' in the definition of relational theories reflects the fact that relational theorists take the relations of apprehension involved in visual experience to be primitive and not explicable in terms of the perceived object being represented as being some particular way (Brewer 2011, p. 94; Soteriou 2011, p. 225).

Some notes. First, this framework characterises theories which aim to give an account of the subjective character of visual experiences. Theories which don't aim to give such an account – such as, perhaps, Burge (2010) – are not covered by this framework. Second, certain representational theories claim to allow visual experience to possess relational elements in virtue of its involving a representational state with singular or *de re* content (Schellenberg 2011): these views sometimes present themselves as being both relational *and* representational. But to the extent that these views take the relations involved in visual experiences to be explained in terms of representational content, they do not count as relational theories according to this framework.

We can use these distinctions to map the three dominant views in the philosophy of perception literature.[5] *Naïve realism* is a relational theory of visual experience: it holds that a certain class of visual experiences, namely those involved in visual perceptions, have a phenomenal character which involves relations of apprehension holding between perceivers and empirical objects and their features. *Sense-datum theories* agree with naïve realism that the phenomenal properties of the experiences involved in visual perception consist of relations of apprehension but deny that the relata of these relations are ordinary empirical objects. *Intentional* theories hold that the phenomenal properties of visual experience are representational.

Intentional theories come in two stripes, depending on whether one takes it that the representational properties which determine the phenomenal character of a visual experience are constrained by the concepts possessed by the subject of the experience. Conceptualist intentional theories hold that subjects of visual experiences must possess the concepts required to specify the content of those

[5] See Martin (2002a), Crane (2005b) and Nudds (2009) for characterisations of the debate along these lines.

 © Gomes 2016

representational states which determine the phenomenal character of visual experience; non-conceptualist intentional theories reject this claim. This terminology should not be confused with the use of the terms 'conceptualism' and 'non-conceptualism' in the Kant literature. In what follows I shall use the terms exclusively in the Kantian sense and put to one side these versions of intentional theories.[6]

Naïve realism is sometimes expressed as the claim that visual perceptions have empirical objects and their features as constituents (Martin 2004, p. 39; Nudds 2009, p. 334; Campbell 2002b, p. 116). This is accurate so long as the notion of constitution is understood to rule out views on which visual perception has an object-dependent representational content (McDowell 1994; Brewer 1999). On these latter views, empirical objects can perhaps be said to constitute visual perception, for they feature in the object-dependent representational content which itself constitutes the phenomenal character of visual perception. This is not the sense of constitution which characterises naïve realism: for the naïve realist, empirical objects constitute visual perception through our non-representationally apprehending them.

A further distinction. We can distinguish *strong* and *weak* forms of each theory. Strong representational theories hold that *all* of the phenomenal properties of visual experiences are representational properties. Weak representational theories hold that at least *some* of the phenomenal properties of visual experiences are representational properties. Mutatis mutandis for relational theories. Some naïve realists explicitly commit themselves to strong relational theories (Brewer 2007, p. 89) but there are no grounds for thinking that all must do so

[6] One question about this framework is where to place the account of perception defended by John McDowell (1994, 1998). Relational theorists often cite him as a proponent of a representational theory of perceptual experience, on the grounds that McDowell takes perception to have a certain sort of content (Brewer 2007). McDowell disputes this characterisation: he takes his view to show that the proper account of the relational aspect of perception cannot do without representational notions and thus that relational views are compatible with perceptual experience having content (McDowell 2008a, 2013a), albeit content which is 'intuitional rather than propositional' (McDowell 2008b, p. 200). This makes it difficult to place McDowell in the above framework. McDowell has some characteristics in common with naïve realist relational theories since he holds that '[i]f something in one's surroundings is in view for one, that is a relation one stands in to it' (2013a, p. 144). But this relational aspect is not primitive since he takes visual perception to be, at root, a representational notion, and to be cashed out in a distinctive notion of content. I will put McDowell's account of perception to one side in what follows. Since McDowell draws explicitly on Kant in setting out his account of perception (2008a, p. 262-4), there is no reason to think that his account is in tension with Kant's views on the nature of perceptual consciousness.

 © Gomes 2016

(Soteriou 2011, p. 225; Phillips 2005). In what follows I will confine myself to strong naïve realism since the early twentieth-century realists who took their naïve realism to be incompatible with Kant's account of perceptual consciousness were strong naïve realists. But to the extent that the conflict between naïve realism and Kant turns on whether there exists a primitive relation of apprehension which is independent of the activity of the understanding, the incompatibility claimed by the early twentieth-century realists will also encompass weak naïve realism even if such views hold that there is more to perceptual consciousness than relations of apprehension.[7]

Why think that naïve realism is incompatible with Kant's account of the involvement of the understanding in perception? The case for incompatibility is set out most forcefully by H.A. Prichard in his 1909 book *Kant's Theory of Knowledge*, written whilst he was a Fellow of Trinity College, Oxford. In his survey of English language commentaries on the first *Critique*, W.H. Walsh sums up the book as follows: '*Kant's Theory of Knowledge* is … not so much a serious study of Kant as a work using Kant as a stalking-horse in an argument for independent philosophical conclusions. Yet it remains, for all the crudities of its interpretation and the insensitivities of its author, an impressive book' (Walsh 1981, p. 732). Both of these comments seem to me to be accurate: Prichard's opposition to Kant takes place in the service of his own philosophical account of perception and knowledge, but the resulting account is one of considerable philosophical interest.

The account of perception and knowledge which Prichard elucidates and defends is one firmly situated in the Oxford Realist tradition. The starting point is the claim that acts of knowing are, by definition, distinct from the things known (Prichard 1909, pp. 108, 118). Knowledge must be sharply distinguished from judgement, opinion and belief: 'Knowledge is *sui generis* and therefore a "theory" of it is impossible. Knowledge is simply knowledge, and any attempt to state it in terms of something else must end in describing something which is not knowledge.' (p. 245) This matters for any account of perception because empirical knowledge presupposes perception (p. 133) and because the perception of spatiotemporal particulars puts one in a position to know about them (pp. 133-5). Perception must therefore be understood as a way of apprehending a reality

[7] See Gomes (2014) for an application of this distinction to the debate about Kant's views on perception.

 © Gomes 2016

independent of our acts of knowing which thereby makes us knowledgeable of it.

Prichard uses this framework to motivate a naïve realist account of perception: perception is the passive faculty by means of which we sensorily apprehend individuals which are independent of the act of perceiving them (pp. 27-30, 133); it is a non-propositional mode of awareness which takes particulars as objects (pp. 44-5); it thus differs fundamentally from the conceptual apprehension of universals or facts (pp. 28, 44). Such an account of perception is required, Prichard claims, if we are to hold on to the Oxford Realist conception of knowledge. And it is this account of perception which forms the basis of Prichard's opposition to Kant.[8]

At issue is Kant's notion of synthesis (1909, p. 217). Kant introduces synthesis as the activity of 'putting different representations together with each other' (A77/B103); it is a 'necessary ingredient' (A120n) in the perception of objects because otherwise the manifold of intuition would be 'dispersed and separate in the mind' (A120); without synthesis we would have only 'unruly heaps' of representations [*Vorstellungen*] (A121). But 'the same function which gives unity to the various ideas in a judgement also gives unity to the mere synthesis of various ideas in an intuition' (A79-80/B105-106) and all combination is an act of the understanding (B130; see also B137). Thus the perceptual presentation of particulars in intuition involves an act of the understanding.

The nature of this act of the understanding is a delicate issue. Béatrice Longuenesse distinguishes 'two aspects of the activity of understanding' (Longuenesse 1998, p. 63): the understanding as rule-

[8] See Kalderon (2011), Kalderon and Travis (2013) and Hatfield (2013) for recent naïve realist readings of Prichard; see also Marion (2000) which suggests a similar reading. Prichard's contemporaries focused on this aspect of his view: see, for example, Price's discussion of Prichard's views on perception in his biographical memoir of Prichard in the *Proceedings of the British Academy* (Price 1947). The Editor has pointed out to me a passage in which Prichard distinguishes perception from 'perception in the full sense' and claims that the latter includes both perception and conception (1909, p. 29). This passage might suggest that Prichard endorses only a weak form of naïve realism since it appears to allow two elements to perceptual consciousness: a primitive relation of apprehension and the conceptual representation of the object perceived. But Prichard does not return to the notion of 'perception in the full sense' in his writings, and his extensive discussion of the problem of illusory experiences does not make use of any conceptual element to perceptual consciousness in providing an account of illusory experiences. For these reasons, I take Prichard's basic notion of perception to capture perceptual consciousness proper. His discussion of perception in the full sense seems like an attempt to capture the fact that there is an intentional use of verbs of sense-perception (as Anscombe 1965 was later to point out). Thanks to the Editor for helpful discussion of Prichard's views.

 © Gomes 2016

giver for the synthesis of the manifold in intuition and the understanding as combiner of concepts in judgement. Similarly, one might distinguish between acts of the understanding which gather representations together in accordance with concepts, and those which involve attributing general attributes to representations. The details don't matter for Prichard: on either construal, perceptual consciousness constitutively depends on the activity of the understanding. And it is this which Prichard takes to be incompatible with naïve realism.

Why is the constitutive claim incompatible with naïve realism? For Prichard, the constitutive claim entails that knowledge 'consists in an activity of the mind by which it combines the manifold of sense on certain principles ... and by which it thereby gives the manifold relation to an object' (p. 230); knowledge-enabling perceptual consciousness thus depends on an act of combination undertaken by 'the productive faculty of the imagination, which, in combining the data into a sensuous image, gives them the unity desired' (p. 225). But according to naïve realism 'the act of combination cannot confer upon [sense-elements] or introduce into them a unity which they do not already possess' (p. 226). The views thus disagree about whether perceptual consciousness relates us immediately to objects without any activity of the understanding. Prichard takes his naïve realism to stand opposed to Kant in renouncing any need for combinatorial activity on the part of the mind to relate us to objects: objects are given to us through primitive, non-conceptual, relations of apprehension.[9]

This criticism is sometimes thought to rely on Prichard's phenomenalist reading of Kant (e.g., Bird 1962, pp. 2-17) but the opposition cuts deeper. It turns on a tension, as Prichard sees it, between the idea of perception as a non-conceptual taking in of empirical particulars and the Kantian claim that perceptual consciousness depends on acts of the understanding. For the naïve realist, the character of our experience is determined solely by a primitive relation of apprehension which holds between subject and empirical particular. For Kant,

[9] See Prichard (1909, ch. 9) for the full development of this argument. Brewer (2006, p. 174f) expresses a related line of thought. Kalderon (2011, p. 236f) and Travis (2013b) discuss Prichard's argument. Prichard's discussion encompasses a number of other issues, including, for example, the question of how to account for the unity of the objects of perceptual consciousness (pp. 224-6); the perceptibility of general features of objects (pp. 226-7); and the grounds for synthesis (pp. 214-7). On some of these topics, Prichard's disagreement with Kant may turn on issues distinct from the question of whether the understanding is constitutively involved in perceptual consciousness. Thanks to a referee for emphasising this point.

 © Gomes 2016

perceptual experience involves a process of synthesis in which sensory elements are combined in accordance with categorial concepts. These claims pull in opposite directions: on one account, perception gets its character from our sensory capacity to apprehend objects; on the other, it arises from the intellectual activity of combining sensory elements in accordance with concepts. Prichard's claim is that it cannot be both. Something has to give.

The same thought finds expression in some contemporary expressions of strong naïve realism. Bill Brewer, Charles Travis and William Fish each oppose their form of naïve realism to any view on which experience has representational content on the grounds that phenomenal character is determined by the empirical objects perceived and not by any intellectual activity of the subject (Brewer 2011; Travis 2013b; Fish 2009). Mark Kalderon makes the opposition explicit: his Prichard-influenced form of naïve realism is presented in opposition to Kant's claims about the role of synthesis in perception (Kalderon 2011, pp. 235-6). As in Oxford Realism, the claim that perceptual consciousness is constituted by primitive relations of apprehension is said to be incompatible with the thought that perceptual consciousness constitutively depends on acts of the understanding.

The response of the Oxford Realists is to reject Kant. But if there are grounds for thinking that Kant's account of sensibility has naïve realist relational elements, then something more needs to be said. In the next section I will provide a brief overview of the reasons for thinking that Kant's account of sensibility can profitably be read on the model of a naïve realist account of visual perception. This leaves the Kantian with a question of how to make compatible naïve realism and Kant's account of the role the understanding plays in perceptual consciousness.

3. Sensibility and the understanding

3.1 Sensibility

Lucy Allais notes that interpretations of Kant 'have gone down many of the same roads as have philosophical accounts of perception' (Allais 2007, p. 465) – to which one might add the clause 'a full decade later'.[10] Kant's account of sensibility has been read on the model of sense-datum theories of perceptual experience (Broad 1978; Strawson

[10] Given that Mike Martin's 'The Transparency of Experience' and John Campbell's *Reference and Consciousness* were both published in 2002, this paper continues the trend.

 © Gomes 2016

1966); conceptualist intentional theories of perceptual experience (McDowell 1998; Abela 2002); and non-conceptualist intentional theories (Hanna 2005, 2008). Allais's own work – despite its reference to contemporary non-conceptual theories (Allais 2009, p. 386) – is best read as completing the taxonomical space by providing support for a naïve realist interpretation of Kant (Allais 2009, p. 394; 2010, p. 60; 2011, pp. 379-83; see also McLear 2016).

What grounds are there for explicating a naïve realist account of visual experience in Kantian phrase? Although the physical sensory capacities of human beings receive little attention in the first *Critique*, Kant makes clear in the *Anthropology* that the five senses are forms of outer sense (7:153), and that vision, in particular, is a form of objective empirical intuition (7:154). Outer sense is a mode of sensibility by which we are presented with objects external to the subject (A22/B37) and, as a form of outer sense, the objects of visual experience are presented to us as situated in space (A22/B37) and distinct from ourselves (*An* 7:156). More generally, outer sense is a mode of our capacity 'to acquire representations [*Vorstellungen*] through the way in which we are affected by objects' (A19/B33) and it does this by means of intuitions [*Anschauungen*] which relate us immediately to the objects of experience (A19/B34). Are there any reasons to think that this account of sensibility involves naïve realist commitments?[11]

An immediate obstacle to any naïve realist interpretation of Kantian sensibility is the frequency with which the term '*Vorstellung*' and its plural '*Vorstellungen*' feature in the first *Critique*. Both Guyer and Wood and Kemp-Smith translate '*Vorstellung*' as 'representation', and Kant explicitly identifies perception as a form of *Vorstellung* in the so-called stepladder (*Stufenleiter*) passage at the start of the Transcendental Dialectic (A320/B376-7). If perception is a form of

[11] Note that we need to distinguish Kant's use of the term 'experience' [*Erfahrung*] from our contemporary use. (See McLear forthcoming, §1 and the next section, below.) Kant is clear that *Erfahrung* involves a process of conceptual synthesis and as such is determined by the concepts a subject possesses (B218-9). This would seem to rule out a naïve realist reading of *Erfahrung*. But *Erfahrung* is 'a kind of cognition requiring the understanding' (Bxvii) and Kant often uses the term to pick out a kind of perceptual judgement rather than a perceptual experiential state (e.g., B166, A176/B218, A189/B234). Since Kant is explicit that visual perception is a form of objective empirical intuition (*An* 7:154; cf. *Prol.* 4:283; *Prol.* 4:283; A180/B222), the question to be pursued here is whether the sensory consciousness involved in outer intuition [*Anschauung*] can be understood on naïve realist lines and Kant's claims concerning the way in which *Erfahrung* is determined by a subject's concepts don't settle that issue. Thanks to a referee for raising this issue.

 © Gomes 2016

Vorstellung, and '*Vorstellung*' is to be translated as 'representation', then it would seem that Kant endorsed some form of representational or intentional theory of perception.

The translation of '*Vorstellung*' as 'representation' is often justified on the grounds that Kant uses the Latin term '*repraesentatio*' in a parenthetical aside which seems designed to explicate the term (A320/B376). But the literal translation of '*Vorstellung*' is a putting (*stellung*) before (*Vor*), and Werner Pluhar's recent translation prefers the term 'presentation', as did Wolfgang Schwarz in his 1982 concise version (Kant 1982) and C.D. Broad in his Cambridge lectures (Broad 1978). More generally there are strong hermeneutical grounds for thinking that Kant's use of the term '*Vorstellung*' doesn't commit him to thinking of perception as representational in the way in which such a claim is understood by intentional theories of perceptual experience.

Let me highlight two issues. The first concerns sensations [*Empfindungen*]. Kant defines a sensation as 'the effect of an object on the capacity for representation, insofar as we are affected by it' (A19/B34), and gives as examples taste and colour, 'effects accidentally added by the particular constitution of the sense organs' as Kemp Smith has it (A28). And, at least with regards to his account of sensation in the first *Critique*, the dominant view has it that Kant took sensations to be non-intentional states: mere modifications of sensory consciousness which do not represent states of affairs.[12] Yet in the stepladder passage, Kant places sensation under the genus *Vorstellung:* a sensation is a *Vorstellung* with consciousness which relates solely to a subject as a modification of its state (A320/B376). If these views are right that sensations are non-intentional states, then Kant's use of the term '*Vorstellung*' cannot commit him to our contemporary notion of representation.

The second issue concerns our understanding of the history of Kant interpretation. It's unsurprising that Broad uses the term 'presentation' to translate '*Vorstellung*' favouring, as he does, a sense-datum interpretation of Kant (Broad 1978, pp. 18-9). For if the term '*Vorstellung*' were to commit Kant to an intentional theory of perception, a relational reading of Kantian experience would be ruled out by fiat – and sense-datum theories are versions of a relational theory. This is sometimes missed by those who treat naïve realism as if it were the

[12] George (1981); Westphal (2004, p. 44); Hanna (2005, p. 254); see Dickerson (2003, p. 24) for an opposing view.

 © Gomes 2016

only form a relational theory can take. And though there is much to object to in sense-datum interpretations of Kant, it is surely too much to hope that their falsity be shown simply by Kant's use of the term '*Vorstellung*'. Charity towards our predecessors requires that we invoke an understanding of '*Vorstellung*' which doesn't settle the issue of whether Kant endorsed a representational or relational theory of perception. The prevalence of '*Vorstellung*' and '*Vorstellungen*' in the *Critique* shouldn't tell against a naïve realist account of Kantian sensibility.

Are there any positive reasons for endorsing such a reading? In the rest of this section I will mention four reasons for thinking that Kant's account of outer sense can profitably be read on the model of a naïve realist account of perceptual experience and that it is therefore appropriate to explicate a naïve realist account of visual experience in Kantian phrase. These reasons are not decisive: it would take more space than I have available to make the case in detail.[13] But I suspect that the strongest consideration against a naïve realist interpretation of Kant is the thought that naïve realism conflicts with Kant's claim that the understanding is involved in perceptual consciousness. To this extent, the wider argument of this paper is intended to buttress the considerations appealed to in this section.

(1) Accuracy conditions: the standard way to explicate what is involved in visual experience possessing representational content is via the notion of an accuracy condition (Tye 2009, p. 252; Siegel 2010; see also Searle 1983, p. 43; Peacocke 1983, p. 5). Visual experiences are representational because they possess conditions which specify ways the world must be in order for the representational content of the experience to be true. Thus if Kant endorsed some form of a representational theory of perception, he would be committed to thinking of perceptual experiences as assessable for truth and falsity.

Yet according to the account of error Kant sets out at the start of the Transcendental Dialectic, the deliverances of the senses do not themselves admit of truth or falsity: 'It is correctly said that the senses do not err; yet not because they always judge correctly, but because they do not judge at all' (A293). The possibility of error enters only with judgement: it is when one judges that things are thus and so that one takes a stand assessable for truth and falsity. But '[i]n a representation

[13] McLear (2016) provides a detailed defence of the ascription to Kant of a form of naïve realism on which perceptual consciousness involves being immediately acquainted with aspects of the mind-independent environment.

 © Gomes 2016

of sense (because it contains no judgement at all) there is no error' (A294/B350). This diagnosis is repeated in the Logic lectures: '*Truth* and *error* are found ... only in judgements' (*BL* 24:83; see also *JL* 9:53, *VL* 24:825). If the deliverances of the senses do not admit of truth or falsity, they cannot set accuracy conditions in the manner required by representational theories of perception.[14]

One response to this line of thought is to distinguish accuracy conditions from truth conditions. One might hold, for example, that only conceptually structured content can be truth-apt whereas non-conceptual content can be assessed for accuracy without possessing truth conditions. And if this is right, then Kant's claim that the deliverances of the senses lack truth conditions is compatible with the claim that they possess accuracy conditions, when accuracy conditions are understood in this weaker way.

The problem with this suggestion is that Kant states, in the passages quoted above, that illusion and error are found only in judgements (A293/B350). And it is hard to see why he would say this if he thought that intuitions had representational content which could be assessed for accuracy. Those who do distinguish accuracy conditions from truth conditions do so because they think that the former can explain illusion and misrepresentation without adverting to the type of conceptually structured content they take to be essential to the latter. Kant's insistence that illusion is to be found only in judgement makes it implausible that he thought the deliverances of the senses were assessable for accuracy but not truth.[15]

(2) Generality: the link between representational content and accuracy conditions holds because states with representational content represent things as being a certain way (Crane 1992, p. 139). Bill Brewer has argued that to represent a particular as being a certain way is to relate that particular to a general condition that must be met in order for the content of the representation to be true (Brewer 2006, pp. 172-4). Generality is thus 'essentially involved in the notion of perceptual content' (Brewer 2006, p. 172). This is clearest in the case of the thought that *a* is *F*: 'Here, a particular object, *a*, is thought to be *a specific general way*, *F*, which such objects may be and which infinitely

[14] McLear (2016, §4.1) develops this line of thought as an argument against ascribing to Kant any form of representational theory.

[15] Note also that there is no evidence that Kant anywhere endorses the claim that intuitions are assessable for accuracy, so there is a burden on a representational theorist who wants to make this move to provide a positive reason for reading Kant in this way. Thanks to a referee for making this point.

 © Gomes 2016

many qualitatively distinct objects are' (Brewer 2006, p. 173, emphasis in original). But it also holds, Brewer claims, for representations with singular or doubly demonstrative content, such as the representation 'that man is thus'. Such representations 'again represent a particular thing as being a determinate general way, which, again, infinitely many qualitatively distinct possible objects are' (2006, p. 173).

What does Brewer mean in claiming that representational content essentially involves generality? I take the thought to be something like the following: all representational content -- whether singular or general – involves specification; it requires a selection from amongst the ways the world could be in order to fix the way which determines the truth conditions for the representational state. This is true even if the selection involves demonstrative terms and is incompatible with the falsity of what is selected. Consider Brewer's example of perceiving a red ball. In this case, the colour and shape are selected as relevant from a set of parameters including the ball's weight, shape and cost. And within the selected parameters of colour and shape, some unique value for its colour and shape are selected as setting the correctness conditions on the experience. Any variation from these unique settings results in the experience being inaccurate. This selection – this singling out – showcases a kind of generality essentially tied to representational content: all content involves the selection and categorisation of determinate general conditions which specify a way the world must be in order for the content of the experience to be true.[16]

This feature of representational content is hard to square with Kant's insistence on the *particularity* of intuitions and his contrast with the *universality* or *generality* of concepts (A19/B33; A68/B93; A320/B376-7; *JL* 9:91). It is concepts which represent 'what is common to several objects' (*JL* 9:91) whereas intuitions singularly and immediately relate us to particular objects without a detour through general marks common to all such objects (A19/B33). Sebastian Gardner speaks for many when he claims that '[t]he distinction of intuition and concept thus corresponds to the distinction between the particular and the general' (Gardner 1999, p. 66; cf. Strawson 1966, pp. 20-1). Thus if outer sense proceeds by means of *Vorstellungen* which lack generality – and if Brewer is right to think that representational theories of perception are committed to

[16] Charles Travis also links the notions of representational content and generality. See the Introduction to his 2013a and many of the essays therein.

 © Gomes 2016

perceptual experiences possessing such generality – then outer sense cannot be modelled on representational theories of perception.[17]

(3) The object-dependence of intuition: outer sense proceeds by means of intuitions, the defining characteristics of which are immediacy and singularity (A19/B33, A68/B93, A320/B377, *JL* 9:91, *VL* 24:905). Kant tells us in the *Prolegomena* that 'an intuition is a representation of the sort which would depend immediately on the presence of an object [*Anschauung ist eine Vorstellung, so wie sie unmittelbar von der Gegenwart des Gegenstandes abhängen würde*]' (*Prol.* 4:281) and later in the second-edition of the *Critique* that an intuition 'is dependent on the existence of the object' (B72). As Allais puts it, 'a clear and straightforward way of understanding [this claim] ... is to see intuitions as representations which essentially involve the presence to consciousness of the particular things they represent' (Allais 2010, p. 59). On this reading, intuitions are object-dependent relations which require the presence of the object intuited.[18]

If this reading is to be rendered plausible, some explanation must be given of those passages in which Kant appears to allow the possibility of intuitive representations in the absence of the objects perceived.[19] But, alongside textual exegesis, defenders of this approach typically offer us further structural reasons for taking intuitions to be object-dependent states. Allais argues that the object-dependence of intuitions is required to make sense of Kant's argument for transcendental idealism in the Transcendental Aesthetic (Allais 2010) and that the object-dependence of intuition plays a crucial role in Kant's account of mathematics (Allais 2010, pp. 61-2). Robert Hanna takes

[17] Can this worry be avoided by arguing that Brewer's notion of generality is wider than Kant's, such that Kant's insistence on the particularity of intuitions is compatible with their involving generality in Brewer's sense? One might argue that since Kant didn't recognise the existence of singular concepts (*JL* 9:91), the particularity of intuitions is compatible with their involving the generality associated with *singular* representational content. But Kant's denial of the existence of singular concepts occurs in the context of his drawing a distinction between the way in which intuitions and concepts relate us to objects, and one way to put the points of Brewer and Travis is that singular content relates us to objects in much the same way as general content, through the setting of determinate general conditions. See Travis (2013a) on the substantive issue and Thompson (1972); Land (2013) for the case against treating Kantian intuitions as singular terms.

[18] See also Hanna (2001, pp. 209-10); cf. Stephenson (2015) for an opposing view.

[19] For example, (B278-9, *An* 7:153; 7:167); see Stephenson (2015) for other relevant passages. One option in accounting for these passages is to claim that they show only that inner intuitions are not object-dependent. See McLear (2014, §6) for a version of this suggestion.

 © Gomes 2016

object-dependence as essential to any understanding of Kant's account of the synthetic (Hanna 2001, ch. 4). Let me add a further consideration.

In the B-Preface, Kant draws a distinction between *cognition* and *thought*. We can have cognition of an object 'only insofar as it is an object of sensible intuition' (Bxxvi; cf. B166) whereas our thoughts 'have an unbounded field' (B166). That is to say, *cognition* of objects requires intuitions; *thought* about objects does not. Compare this to the characterisation of cognition and thought offered in the footnote at Bxxvi:

> To **cognize** an object, it is required that I be able to prove its possibility (whether by the testimony of experience from its actuality or *a priori* through reason). But I can **think** whatever I like, as long as I do not contradict myself, i.e., as long as my concept is a possible thought, even if I cannot give assurance whether or not there is a corresponding object somewhere within the sum total of all possibilities. (Bxxvi, emphasis in original)

Here the claim is that *cognition* of objects requires being able to prove the possibility of the object in question; *thought* about objects does not.

How do these two characterisations of the distinction between cognition and thought relate to one another? Kant seems to think that cognition requires sensible intuition because intuition enables subjects to prove the possibility of the objects intuited. But why should he think that? The object-dependence of intuition provides a straightforward answer: intuitions allow us to prove the possibility of objects because such intuitions are not possible in the absence of the object perceived. When a subject intuits a particular object she stands in a relation which requires the presence of that very object: it is for this reason that intuition enables you to prove an object's possibility. Taking intuitions to be object-dependent states explains and illuminates Kant's account of the distinction between cognition and thought.[20]

Such object-dependence can be easily modelled on a naïve realist relational account of perception. According to such accounts, the objects of perception feature as constituents of a subject's perceptual experiences. It follows from this claim that perceptual experiences are object-dependent: having an experience of this kind requires the

[20] See Gomes and Stephenson (forthcoming) for further discussion of the relation between the object-dependence of intuitions and the possibility of cognizing objects.

 © Gomes 2016

presence of the object perceived. A naïve realist account of outer sense which takes intuitions to be relations to objects captures the object-dependence of Kantian intuitions.

As is stands, this consideration is compatible with a representational theory of perceptual experience which takes the content of perceptual experience to be object-dependent. But note that it doesn't follow from the fact that the content of a state is object-dependent that the state itself is object-dependent (Martin 2002b) so more would need to be said about how the object-dependence of the content of experience secures the object-dependence of intuitions themselves. More importantly, Kant is clear that, in the case of thought, the fact that the content of a thought is non-contradictory does not suffice to show that its content is *really possible* (Bxxxvi). The onus is thus on the representational theorist to explain how placing object-dependent content within perceptual experience suffices to show that the content of that experience is really possible. For if the non-contradictory content of a thought is insufficient to prove its possibility, it is unclear how making that content part of a perceptual state serves to make it sufficient.

(4) The Fourth Paralogism and the Refutation of Idealism: the Kantian phrase with which Strawson summarises his account of visual experience alludes to Kant's claim in the Refutation of Idealism that 'the consciousness of my own existence is at the same time an immediate consciousness of the existence of other things outside me [*das Bewußtsein meines eigenen Daseins ist zugleich ein unmittelbares Bewußtsein des Daseins anderer Dinge außer mir*]' (B276). Kant's target in the Refutation is the 'problematic idealism' of Descartes, a topic which was originally covered in the A-Edition of the Fourth Paralogism. And there is much in both sections which lends support to a naïve realist account of Kantian sensibility.

Consider first the details of Kant's attack on problematic idealism. Kant's charge in these passages is that problematic idealism leaves the existence of material objects open to question since it requires one to infer the existence of material objects as the cause of representations which lie within us (A368-372; B276). This complaint has both epistemological and cognitive dimensions. Epistemological because such idealism leaves the existence of material objects 'far from certain' (A372) and thus unknown; cognitive because the inference from effect to cause leaves it 'absolutely impossible to comprehend how we are to acquire cognition [*Erkenntnis*] of their reality outside us' (A378). In both cases, the causal story by means of which problematic

 © Gomes 2016

idealism aims to explain our cognitive contact with the world is claimed to be lacking.

This charge reappears in the naïve realism literature. John Campbell writes of views which take the relation between perceptual experiences and empirical objects to be purely causal: 'How can effects provide you, the subject, with any conception of what their causes are like?' (Campbell 2002a, p. 132; cf. Child 1994, pp. 147-9). Campbell's target here is not only sense-datum models of perceptual experience, but also representational theories which hold that material objects causally determine the representational content of experience (Campbell 2002a, pp. 132-3). The thought is that such theories leave us cognitively isolated from the world precisely because they involve a 'conception of conscious experience as something that is merely an effect of external objects' (Campbell 2002a, p. 133). That is, they assume 'that the only immediate experience is inner experience, and that from that outer things could only be **inferred** …' (B276, emphasis in original).[21]

Campbell uses this argument to motivate a naïve realist relational account of perception. On a naïve realist view, '[w]e have to regard experience of the object as reaching all the way to the object itself': only so can we allow conscious experience a role in explaining our ability to think and know about material objects (Campbell 2002a, p. 143). Similarly, Kant takes the upshot of his argument against problematic idealism to be a view of perceptual experience on which 'outer experience is really immediate' (B276), one which 'grants to matter, as appearance, a reality which need not be inferred but is immediately perceived' (A371). It is only by thinking of perceptual experience as reaching out to the particulars in our environment that we can 'prove the existence of objects in space outside me' (B275). The Refutation of Idealism supports a naïve realist reading of outer sense.

These four considerations, taken together, do not force upon us a relational reading of Kantian sensibility, but it is not obvious that we should expect Kant's texts to be unequivocal in this way. Instead we should note how these aspects of Kant's account of sensibility can be easily accommodated within a relational framework. I take this to provide some support for the explication of a naïve realist account of visual experience in Kantian phrase.[22]

[21] See Gomes (2013) for more on the relation between this aspect of naïve realism and Kant. Cook Wilson makes a similar point in his 1904 letter to G.F. Stout (Cook Wilson 1926, pp. 769-73).

[22] Is the naïve realist reading of Kantian sensibility in tension with Kant's scientific realism? Kant draws a distinction between the empirical thing-in-itself and the empirical appearance at

 © Gomes 2016

3.2 Understanding

What are our options? I've set out the Oxford Realist case for thinking that naïve realism is incompatible with the claim that acts of the understanding are constitutive of perceptual consciousness. And I've set out the reasons for thinking that Kant's account of sensibility has naïve realist elements. One way to resolve this tension is to deny that Kant thought the understanding constitutive of perceptual consciousness. There are two ways one might do so. First, one might deny that the understanding is constitutive of perceptual consciousness but hold that Kant took it to be necessarily involved in a certain sort of perceptual consciousness, namely that enjoyed by rational beings with discursive intellects. Second, one might take the stronger line that the understanding is not constitutive of perceptual consciousness nor even necessarily involved in the perceptual consciousness of rational beings with discursive intellects. It is important to recognise that these are two different ways of denying the constitutive claim; the understanding may be necessarily involved in the perceptual consciousness of rational beings without being constitutive of it.

This gives us three options when thinking about the role of the understanding in perceptual consciousness: acts of the understanding are constitutive of perceptual consciousness; acts of the understanding are necessarily involved in the perceptual consciousness of rational beings but not constitutive of perceptual consciousness; acts of the understanding are not constitutive of perceptual consciousness nor necessarily involved in the perceptual consciousness of rational

---

A45-6/B62-3, and it is natural to read this as a form of the primary/ secondary quality distinction. More generally, for Kant, the properties that the empirical thing-in-itself really possesses are those which science ascribes to it – roughly, force and motion – whereas the properties possessed by the empirical appearance are dependent, in some way, on human perceivers and their sense organs. Does this mean that empirical objects don't have the qualities that we naïvely attribute to them, contrary to naïve realism? First, note that naïve realism can allow that 'the idiosyncracies of the perceiver may affect phenomenal content' (Campbell 2002b, p. 119) since perceptual consciousness is constituted by relations, and the nature of both relata can affect the nature of the perceptual relation. And second, naïve realism is incompatible with scientific realism only if it follows from scientific realism that the things we ordinarily take ourselves to be related to don't actually exist – that the world is nothing more than force and motion. And there is no reason to think that Kant's distinction between the empirical thing-in-itself and empirical appearance commits him to thinking of the empirical appearance as less than empirically real. See Langton (1998, ch. 7) for a careful discussion of Kant's scientific realism, and Hanna (2006, ch. 4) for an argument that this scientific realism doesn't entail the idealism of manifest objects and their qualities. Thanks to a referee for raising this issue.

 © Gomes 2016

beings. Call these the strong, intermediate and weak claims respectively.[23]

Recent *non-conceptualist* interpretations of Kant defend the weak claim. According to non-conceptualist readings, Kant's view is that we can be perceptually presented with particulars without any input from the active faculty of the understanding (Allais 2009, 2012; Hanna 2001, 2005; McLear 2016). Since Kant introduces the understanding as 'a **faculty for judging**' and tells us that all judgement proceeds via concepts (A69/B94, emphasis in original), this is often expressed as the claim that one can be perceptually presented with particulars without the application of concepts. Non-conceptualist readings hold that 'for Kant, the application of concepts is not necessary for our being perceptually presented with outer particulars' (Allais 2009, p. 394).

Such readings can be used to eschew any felt tension between Kant's theory of cognition and a naïve realist account of visual experience. For if Kant thinks that we can be visually presented with an outer particular without applying any concepts, then there is no obstacle to our conceiving of such visual presentation as consisting in a relation of apprehension holding between subject and perceived particular. By denying that the understanding is necessarily involved in perceptual consciousness, non-conceptualism opens up the space for a Kantian form of naïve realism. Lucy Allais (2009, 2012) motivates the link between non-conceptualism and relational theories of perception in just this way.

Some *conceptualist* interpretations adopt the strong claim. Griffith, for example, claims that, for Kant, 'perception depends on empirical synthesis, which depends on a transcendental synthesis according to the categories' (2010, p. 22), where this dependence is to be understood as '(conceptual/transcendental) dependence or grounding' (p. 20). It is claims of this sort which the Oxford Realists took to be

[23] Defenders of the weak claim can allow that there are some cases in which acts of the understanding cognitively penetrate perceptual consciousness. One example might be the kinds of perceptual experiences involved when subjects with sophisticated conceptual repertoires engage with certain works of art. This means that there is a spectrum of views one can hold, with the intermediate claim that the understanding is necessarily involved in, though not constitutive of, the perceptual consciousness of rational beings lying at one end, and the view that the understanding is not involved in perceptual consciousness at all lying at the other. Extant weak views seem to fall towards the latter extreme. Both Allais and Hanna make much of the fact that the understanding is involved only in the capacity to think about empirical objects (Allais 2012, pp. 103-6) and make objectively valid judgements (Hanna 2013), and neither discusses any cases in which the understanding penetrates perceptual consciousness. But one could endorse a weak view which allowed some cases of penetration. Thanks to a referee for discussion.

 © Gomes 2016

incompatible with naïve realism and ascribing such a view to Kant is in tension with the thought that his account of sensibility involves naïve realist elements.

The intermediate claim falls in between the standard characterisations of conceptualism and non-conceptualism in the Kant literature. Hannah Ginsborg, for instance, defends conceptualism by arguing that 'Kant is committed to a view on which perceptual synthesis involves the exercise of understanding' (2007, p. 11) – a claim which is neutral between the intermediate and strong formulations. On the intermediate view, the understanding is necessarily involved in the perceptual consciousness of rational beings but it is not constitutive of perceptual consciousness. This is a non-conceptualist position, in so far as it denies that the understanding is constitutive of perceptual consciousness. But it is a conceptualist position, in so far as it takes the role of the understanding, in discursive beings, to extend beyond mere thought about objects into all instances of perception.

The Oxford Realists claim that naïve realism is incompatible with the strong claim. But they take it to follow from this that the understanding is not involved in perceptual consciousness at all. Similarly, non-conceptualist readings that ascribe a form of naïve realism to Kant deny that the understanding is necessarily involved in the perceptual consciousness of discursive beings. The availability of the intermediate reading suggests a middle position. It may be possible to defend a Kantian form of naïve realism on which the understanding is necessarily involved in discursive perceptual consciousness without being constitutive of perceptual consciousness. This would make naïve realism compatible with Kant's claims about the role of the understanding in perceptual consciousness without endorsing the weak reading. It is an intermediate position of this sort that I offer below.[24]

Before I do so, it will be useful to consider some of the reasons for rejecting the weak reading. We can think of the debate between the strong and intermediate views on one side, and the weak view on the other, as a dispute about where best to limn the domain of the understanding. Strong and intermediate readings hold that the

[24] McDowell may seem to offer us an intermediate position of this sort, for although he takes the understanding to be constitutive of perceptual consciousness, he is clear that he means it to be constitutive of *our* perceptual consciousness, allowing that there might be other forms of perceptual consciousness which stand to our form of perceptual consciousness as a different species of the same genus (McDowell 2011, pp. 20-1, 54-7). Matt Boyle (2012) provides a helpful framework for thinking about this kind of position. I comment on the relation between my proposal and that of McDowell and Boyle below.

 © Gomes 2016

understanding reaches out to perception itself: the understanding is necessarily involved in the perceptual consciousness of discursive beings. Weak readings deny that the understanding is required for perceptual consciousness: the perceptual presentation of particulars can take place in the absence of concepts, but we require the understanding in order to cognize them in a certain way. On this view, the involvement of the understanding may still be required for us to engage in a certain sort of thought about objects but it is not required for the perceptual presentation of particulars.

How should we decide between these readings? A multitude of considerations have been raised by parties on both sides of this debate and I do not have space to consider them all here. But it will be useful to begin by summarising some of the reasons given for rejecting the weak reading's claim that the understanding is not necessarily involved in our being perceptually presented with outer particulars.[25]

The falsity of the weak reading is sometimes motivated by appeal to the opening paragraphs of the Transcendental Logic in which Kant distinguishes sensibility from the understanding and notes their interrelation (A50-52/B74-76). But this will not suffice. Kant's claim in these passages is only that the unification of sensibility and the understanding is required for *cognition* [*Erkenntnis*] (A51/B76), and this falls importantly short of claiming that their co-operation is required for perception itself.

One might attempt to draw a link from these passages to perception by noting that Kant held experience [*Erfahrung*] to be 'a kind of cognition requiring the understanding' (Bxvii), for if the understanding is active in cognition, and if experience is a kind of cognition, then the understanding is active in experience. But Kant's use of the term 'experience' is not continuous with that of contemporary philosophers of perception and it is open for non-conceptualists to hold that at least some of Kant's uses of the term pick out a form of judgement made on the basis of perceptual experience rather than the experience itself (e.g., B166, A176/B218, A189/B234). On this reading, those passages in which Kant claims that *Erfahrung* requires the active, combinatorial input of the understanding (A93/B126) show only that the understanding is required for a certain sort of empirical judgement or thought.

[25] I provide a more detailed discussion of the debate in Gomes (2014) from which some of the following is drawn.

 © Gomes 2016

There is nothing thus far which motivates the involvement of the understanding in *perception*.

It is for this reason that the debate between weak readings and their opponents has largely focused on whether the understanding is active in intuition [*Anschauung*]. In contrast to his use of term '*Erfahrung*', Kant tells us explicitly that visual perception is a form of objective empirical intuition (*An* 7:154; cf. *Prol.* 4:283) and that empirical intuition is the means by which we are perceptually presented with objects (A180/B222; *Prol.* 4:283). Thus if the understanding is essentially involved in *Anschauung* itself, this would seem to tell against the weak claim since it would show that the discursive activity of the understanding is involved in the very perception of distinct particulars.

There are two prominent reasons why commentators have thought that Kant held the understanding to be involved in intuition: his account of the role synthesis plays in the representation of intuitions and his aims in the Transcendental Deduction of the Categories. Let me summarise these considerations briefly.[26]

The central points on synthesis were discussed above. Kant takes perception to involve acts of synthesis, ones in which the manifold of intuition, given to us through sensibility, is taken up and combined through an act of the understanding. This act of the understanding – one in which 'the understanding determines the sensibility' (B160n.) – confers a unity on the manifold of intuition. It does so by combining the manifold of intuition in accordance with the a priori rules which govern synthesis. These rules are the categories. 'Consequently all synthesis, through which even perception itself becomes possible, stands under the categories…' (B161). The understanding is necessarily involved in perceptual consciousness in virtue of the perceptual presentation of particulars requiring an act of synthesis which combines the manifold of intuition in accordance with the categories.

Second, Kant's stated aim in the Transcendental Deduction is to show that 'without their [the categories'] presupposition nothing is possible as object of experience' (A93/B125) for 'the objective validity of the categories, as *a priori* concepts, rests on the fact that through them alone is experience possible' (A93/B126). His argument in support of this conclusion turns on the necessity of combining the

[26] I discuss these issues at greater length in Gomes (2014), arguing that weak readings cannot provide an account of the Transcendental Deduction's role in answering Humean scepticism about our justified use of a priori concepts.

 © Gomes 2016

manifold of intuition. Since 'the combination of the manifold in general can never come to us through the senses' (B129), Kant argues that it must be the understanding, through the use of a priori concepts, which brings representations under the original synthetic unity of apperception required for combination. The conclusion is that 'the manifold in a given intuition also necessarily stands under the categories' (B143).[27]

These considerations have been challenged and disputed by those who advocate a weak reading. I will not assess their responses here. For our purposes it is enough to note the prima facie textual evidence for taking Kant to hold that the understanding is necessarily involved in perceptual consciousness, contrary to the weak reading.

Do these considerations motivate the strong reading? In the final section of this paper I will return to the considerations noted above to see whether they can be met by a view which holds that the understanding is not constitutive of perceptual consciousness but is, nevertheless, necessarily involved in the perceptual consciousness of rational beings. For now I want to show that there is nothing which precludes a Kantian form of naïve realism from allowing the necessary involvement of the understanding in the perceptual consciousness of rational beings, contrary to the implication of the Oxford Realists. This is the task to which I turn in the next section.

4. Unification

Let us return to Oxford Realism. Prichard's opposition to Kant's account of perceptual experience centres on an assumption about the character of perceptual consciousness: that it either involves a primitive relation of apprehension or a synthesis of sensory elements in accordance with concepts. Part of what is valuable about Prichard's discussion is that he teases out the way in which these options are distinct: the primitive relation of apprehension which characterises naïve realism is fundamentally different in kind from the state with categorial content which arises from Kant's synthesis of sensory elements. And since they are fundamentally different, any aspect of phenomenal character which is explained by relations of apprehension

[27] Note that these considerations motivate ascribing to Kant the specific claim that certain a priori concepts – the categories – are essentially involved in perceptual consciousness. In what follows I will take it that the understanding is necessarily involved in perceptual consciousness for Kant through the activation of categorial concepts specifically.

 © Gomes 2016

cannot involve the synthesis of sensory elements in accordance with concepts. But, given this insight, why should we accept that there is no more to perceptual consciousness than primitive relations of apprehension? In short: why can perceptual consciousness not have both aspects?

This question is raised with particular force by Prichard's successor at Trinity College, H.H. Price, in his 1932 book *Perception*. Price was an undergraduate in Oxford and a post-graduate in Cambridge, and his book bears the imprint both of Cook Wilson's Oxford Realism and the Cambridge tradition of G.E. Moore and C.D. Broad. In *Perception* Price defends a sense-datum theory of perception according to which sensory experience involves being aware of a 'given' object of experience, which Price terms a 'sense-datum' (1932, p. 3). This involves one important point of contact with and one important divergence from the naïve realism endorsed by Prichard. The point of contact is that both share an understanding of perceptual experience as involving a primitive and non-conceptual relation of sensory apprehension: in the terminology of §2, both views endorse *relational* theories of perception. The divergence is that Price, in common with Moore and Broad, takes the domain of this relation to be wider than that of mere perception: all sense experience, including illusions and hallucinations, involves the sensory apprehension of an object of experience. Perception cannot be, then, a relation to ordinary, empirical objects.

For our purposes, what is important about Price's discussion is that he combines this relational account of sensory apprehension with a conceptual aspect to perceptual consciousness. Perceptual acceptance – the most basic form of perceptual consciousness (p. 139) – has two elements: the sensory apprehension of a non-physical sense-datum and the *being under an impression that* things are a certain way. This latter element, which Price borrows from Cook Wilson (p. 140), is a representational state in which the subject takes for granted *that* things are a certain way. It is part of perceptual consciousness proper, and not simply a postliminary judgement based on the sensory acquaintance (pp. 140-1): when one perceives a tree, one is acquainted with a non-physical sense-datum and under the impression *that* it is a tree: 'The sense-datum is presented to us, and the tree dawns on us, all in one moment' (p. 141).

On Price's account, perceptual consciousness thus has two elements: a relational aspect of sensory apprehension, the objects of which are particulars; and an intentional aspect of taking for granted that things are a certain way, the object of which is a proposition. Taking

 © Gomes 2016

for granted that things are a certain way does not involve judging that things are that way – it is pseudo-intuitive rather than discursive (p. 150) – but the content which one takes for granted is of a sort to feature in judgement (p. 167). Thus perceptual consciousness involves both a non-conceptual relation to non-physical particulars and the conceptual representation of things as being a particular way. Contrary to Prichard's assumption, perception has both relational and representational aspects.[28]

Price's one-time student, Wilfrid Sellars, follows him in endorsing a dual component theory of perception, one which is explicitly Kantian in motivation and form (1956, 1968, 1978). According to Sellars, perceptual consciousness has two aspects: an intentional aspect of taking things to be a certain way and a sensory aspect of sensing. The former draws on the conceptual capacities of the subject whereas the latter is wholly non-conceptual. And whilst there is some dispute amongst readers of Sellars as to how we are to understand this sensory component, at the least it seems clear that whereas Price took the sensory component to consist of a primitive relation of apprehension to sense-data, Sellars takes it to involve some sort of sensory modification of the subject. The aim of Sellars's dual component account is to respect Kant's distinction between sensibility and the understanding whilst showing how the latter can be necessarily involved in the perceptual consciousness of rational beings. The result is a self-consciously Kantian version of a dual component theory.

The key challenge for such theories is to explain how the two components of perceptual consciousness relate to each other such that we have one unitary, though complex, perceptual state. Sellars takes the relation between the two components to be causal (1963, pp. 90-1) but this raises a question about whether the intentional component is really part of perceptual consciousness proper. Price denies that the non-conceptual component is temporally prior to the intentional component (1932, pp. 150-1) but doesn't explain what binds them together in perceptual consciousness. This is a topic that any dual component theory will need to address.

Nevertheless, we have a model for making naïve realism compatible with the involvement of the understanding in perceptual

[28] It is unclear to me whether Price takes both of these elements to contribute towards the phenomenal character of perceptual consciousness. If he does, then he should be classified as a *weak relational* theorist in the terminology of §2. If not, his account is close to the one I offer below.

 © Gomes 2016

consciousness. We need to recognise, with Price and Sellars but contra Prichard, that perceptual consciousness can have both relational and representational aspects. But instead of taking the non-conceptual aspect of perceptual consciousness to be a relation to sense-data, as in Price, or mere sensory modification of the subject, as in Sellars, we need to insist that the phenomenal properties of perceptual consciousness are constituted by relations of apprehension to ordinary empirical objects. The result would be a naïve realist account of visual experience in Kantian phrase: visual perception involves a primitive relation of apprehension which determines the phenomenal character of the visual experience and intentional content which represents empirical particulars as falling under the categories.

Price himself hints at just such a Kantian position, noting that perceptual consciousness would involve, on his view, concepts 'which are not exemplified in the intuited data at all and cannot be abstracted from them - namely those concepts which make up the notion of "material thinghood"' (Price 1932, pp. 168-9). By replacing Price's sense-data with the ordinary, empirical particulars of Prichard's discussion, we have the outlines for a Kantian form of naïve realism on which one is immediately related to empirical particulars and under the impression that they are subject to categorial concepts. The rest of this section expands on this suggestion.

The picture I will suggest looks as follows: Kantian intuitions are non-conceptual apprehensions of empirical particulars, distinct from ourselves and situated about us in space. These relations alone determine the phenomenal character of visual experience: hence *naïve realism*. But when one stands in such a relation to empirical particulars, the manifold of intuition is synthesised in accordance with the categories in a process governed by the understanding. The result is a representational state which represents empirical particulars *as* subject to the categories. This representational state is not itself an intuition; it is the representation of an intuition and its representational properties are not among the phenomenal properties of the visual experience. But the obtaining of the state bears a necessary connection to the presence of the intuition: a representational state with this content obtains in virtue of a subject with a discursive intellect standing in an intuitive relation of this kind. The understanding is thus necessarily involved in perceptual consciousness, not through constituting the intuition itself, but through the presence of a representational state

 © Gomes 2016

which obtains whenever a discursive intellect intuits empirical particulars in a certain way.[29]

On this account, visual experience has two elements: intuitions which determine the phenomenal properties of visual perceptions and a representational state which obtains partially in virtue of those intuitions. We can mark these two aspects as follows: when a subject S visually perceives an ordinary empirical object E,

(1) S intuits E, and

(2) It seems to S as if she intuits E.

The relation in (1) is a non-conceptual relation of apprehension; the representational state in (2) is the outcome of a process of synthesis and it represents E as subject to the categories. The two components are related: the representational state in (2) obtains *in virtue of* a subject with a discursive intellect standing in the non-conceptual relation of apprehension.

Let me expand on these claims in turn:

(1) *Intuitions are non-conceptual apprehensions of empirical particulars:* this claim secures the naïve realist account of sensibility motivated in §3. It provides a straightforward reading of Kant's claim that intuitions relate us to objects singularly and immediately. Intuitions are singular because they are relations to particulars (A320/B377) and immediate because they present us with particulars directly (cf. Brewer 2006, pp. 172-3). This explains why 'all thought ... must ultimately be related to intuitions, thus, in our case, to sensibility, since there is no other way in which objects can be given to us' (A19/B33): intuitions are non-conceptual apprehensions of empirical particulars which give us objects in such a way that they can be the objects of thought.

(2) When one stands in such a relation to empirical particulars, the manifold of intuition is synthesised in accordance with the categories in a process governed by the understanding: although intuitions are non-conceptual, they still have structure. This is how they differ from sensations: non-intentional modifications of a subject's consciousness (A320/B376). The matter

[29] This way of thinking about the phenomenal character of perception draws on Matthew Soteriou's 'Content and the Stream of Consciousness' (2007) and, in particular, his account of how mental processes relate to mental states; see his (2013) for further development and detail.

 © Gomes 2016

of empirical intuition – its manifold – is sensation; its form is space, the pure form of outer intuition. It is the matter of intuition – sensation – which is open to synthesis by the imagination in accordance with the categories.

(3) *The result is a representational state which represents empirical particulars as subject to the categories:* the synthesis of the manifold of intuition is a rule governed process, guided by the understanding. As outlined in §§15-19 of the B Deduction, the role of this synthesis is to take the sensory elements which are given to us in sensibility and bring them under the unity of apperception. In order for it so to bring them, they need to be synthesised in accordance with the a priori rules of combination. The result is a representational state which represents the world *as* a world of persisting unities, subject to causal relations.

Since this synthesis is governed by the a priori rules of combination, the only concepts which are guaranteed to feature in the representational state are the a priori concepts which characterise the perceived particular. The fact that the intuitive apprehension of the particular can be characterised by certain *empirical* concepts makes available the possibility of those concepts featuring in the representational state, perhaps through rules of association understood as operating in a largely Humean manner. But the involvement of empirical concepts is not a necessary feature of the state.

(4) *This representational state is not itself an intuition; it is the representation of an intuition:* Prichard's claim is that those primitive relations of apprehension which constitute the character of perceptual consciousness cannot themselves be the outcome of a product of synthesis. This is because those primitive relations of apprehension relate us immediately to objects without any input from the understanding, whereas synthesis is undertaken by the understanding in accordance with concepts. It follows from this that the synthesis of the manifold of intuition cannot result in an intuition: synthesis results in a representational state with categorial content whilst intuitions are non-conceptual relations to particulars. One might think this to be a juncture at which this account is *Kantian* rather than Kant's own, for Kant's account of

 © Gomes 2016

synthesis is often read as holding that the product of synthesis is an empirical intuition (Griffith 2012, §5). I return to this issue in §5. For now it is enough to show that this Kantian form of naïve realism can allow the necessary involvement of the understanding in our perceptual consciousness.

Note, though, that this does involve an important departure from the account of perceptual consciousness offered by Price. In Price's model, the intentional aspect of perception represents how things are in the world; its content concerns only the empirical particular and its perceived properties. In contrast, the representational state which results from synthesis is one which represents the intuition, i.e., the non-conceptual relation of apprehension which a subject stands in to the relevant empirical particular. Experience thus has an intentional content which also involves both the empirical particular and the *subject* of the experience: it seems to the subject as if *she* intuits the empirical particular. I will return to the reasons for this divergence in §5 below.

(5) *... and its representational properties are not among the phenomenal properties of the visual experience:* since visual perception is a form of empirical intuition (*An* 7:154; *Prol.* 4:283), it is intuitions which determine the phenomenal character of visual experience. The representational state with categorial content which is the outcome of synthesis is not an intuition, so its properties are not amongst the phenomenal properties of visual experience.[30] Nevertheless, there is a tight connection, to be outlined below, between the intuitions which determine the phenomenal properties of visual experience and the representational state which arises through categorial synthesis of the manifold of intuition. When a subject with a discursive intellect has an experience of a certain sort – when she has intuitions of kind *I* – she stands in representational state *R* as a result of the combinatory activity of the understanding. This connection is necessary. Since *R* obtains partially in virtue of *I*, one can specify the phenomenal character of the visual experience indirectly by highlighting state *R*, to which *I* has given rise. This means that a specification of the phenomenal character

[30] This is a requirement of strong naïve realism. Removing the restriction would result in a form of weak naïve realism. See §2 above.

 © Gomes 2016

can appeal to state *R* despite it being *I* which determines that phenomenal character (cf. Soteriou 2007, pp. 557-9).

(6) *... the obtaining of the state bears a necessary connection to the presence of the intuition:* there is a necessary connection between the presence of an intuition in a subject with a discursive intellect and the obtaining of the relevant representational state. This connection is not causal: the presence of the intuition does not cause a process of synthesis which gives rise to the representational state. Rather, the representational state obtains *in virtue of* a subject with a discursive intellect standing in the intuitive relation.

This relation of *obtaining in virtue of* is a distinctive relation of metaphysical dependence. It has been the subject of recent discussion in metaphysics with a number of metaphysicians arguing that we need to make use of such a notion in formulating metaphysical disputes (Fine 2001; Rosen 2010). Some examples standardly used to illustrate this relation include the claims that an act is pious *because* the gods love it; that the mental is *nothing more than* the physical; that legal facts are *determined by* the social facts. Matt Soteriou (2007, pp. 551f) uses an example of Helen Steward's (1997, p. 72f) to illustrate the connection in question: the state of a body of liquid being at a certain temperature for an interval of time *obtains in virtue of* the motion of its molecules. In each case, the higher level item is grounded in the existence of the lower level item providing us with 'a distinctive kind of metaphysical explanation, in which explanans and explanandum are connected, not through some sort of causal mechanism, but through some constitutive form of determination' (Fine 2012, p. 37).

Recent discussions have attempted to characterise the logical features of this relation. Two of these are relevant for our discussion. First, this grounding relation is stronger than a mere modal connection, since two items can necessarily covary without one holding in virtue of the other (Dancy 1981, pp. 380-2; Fine 2012, p. 38). It is for this reason that the relation needs distinguishing from mere supervenience. Second, the grounded item is not identical with its ground, nor is it any less real than its ground: both ground and grounded exist, and the latter obtains in virtue of the former (Fine 2012, pp. 38-9). Some will wonder whether there is any such relation – stronger than mere causation, stronger than necessary covariance, weaker than

 © Gomes 2016

identity – which can do this work, but I will proceed as if the relation is unproblematic.[31]

My claim is that the representation of the intuition obtains in virtue of the intuition being present in a subject with a discursive intellect.[32] The intuition is not identical to the representational state, because the intuition is a non-conceptual relation of apprehension whilst the representational state involves categorial concepts, and intuitions cannot be identical to or constitute conceptual states (*JL* 9:91). Rather the representation of the intuition obtains partially *in virtue of* the presence of that intuition, for *what it is to be* in that representational state just is for the intuition to be present in a subject with a discursive intellect. The explanation of this dependence is that the manifold of intuition is necessarily subject to an act of categorial synthesis in order to bring the manifold under the unity of apperception. Since this synthesis is guided by the a priori rules of the understanding, the resultant representational state is one which makes use of categorial concepts. On this way of thinking about the mechanics of Kant's theory of perceptual experience, the role of synthesis is to import the intuition into the content of the representational state whilst explaining why it must be that such states obtain whenever discursive intellects stand in intuitional relations.

It is this relation of metaphysical dependence – this relation of *obtaining in virtue of* – which explains why the two elements should be classed as parts of one complex perceptual state. For the two elements are not merely causally connected, as in Sellars's dual component theory, nor merely co-present, as perhaps in Price. Rather, the intentional element is grounded in the relational element and the result is a

[31] This seems acceptable in a Kantian context since we plausibly need to appeal to something like a grounding or in virtue of relation in order to characterise a number of Kant's own views. Consider, for instance, his claim that real possibilities are grounded in how things are in actuality (*OPA* 2:77-79). For further discussion of the grounding relation see the papers in Correia and Schnieder (2012).

[32] For the sake of convenience, let the grounding relation hold between facts, where the right hand side of the relation encompasses a plurality (Audi 2012, Rosen 2010). And let the use of square brackets pick out facts. Then we can state the claim as follows: [it seems to S as if she intuits E] is grounded in [S intuits E], [S has a discursive intellect]. If we further distinguish full grounding from partial grounding [Fine 2012, p. 50], then [it seems to S as if she intuits E] is fully grounded in [S intuits E], [S has a discursive intellect]; and partially grounded in [S intuits E]. Since only full grounding suffices for metaphysical necessitation, only [S intuits E], [S has a discursive intellect] will necessitate [it seems to S as if she intuits E]. And since the grounding facts are only sufficient for the grounded fact, the proposal is neutral on the question of whether [it seems to S as if she intuits E] can obtain without [S intuits E], [S has a discursive intellect].

 © Gomes 2016

complex episode of perceptual consciousness with both relational and intentional aspects.

A final point. Note that this connection between the representational state and the intuition holds only for subjects with a discursive intellect. This allows for the possibility of creatures who lack a faculty of understanding standing in intuitive relations to empirical particulars without the obtaining of a representational state involving categorial concepts. Some have suggested that there are grounds for holding that Kant thought of non-human animals in this way (Allais 2009, pp. 406-7; McLear 2011). This proposal is compatible with such a view: non-human animals are perceptually related to empirical particulars in the environment despite lacking the capacity to represent such particulars as falling under concepts. As Kant puts it in his discussion of the different levels of cognition in his lectures on logic, '[a]nimals are acquainted with objects too, but they do not cognize them' (*JL* 9:64-5; see also *VL* 24:846; FS 2:59) and this is precisely because they do not represent them as subject to concepts.[33]

This model of visual experience combines a naïve realist account of the phenomenal character of visual experience with the necessary involvement of the understanding in perceptual consciousness. It thus provides an account of visual experience intermediate between the claim that the understanding is constitutive of perceptual consciousness and the claim that the understanding need not be involved in perceptual consciousness at all, doing justice to those strands of Kant's

[33] If non-human animal perception lacks a representational component, does that mean that animals lack the capacity to perceive things *as* being a certain way? Many think it plausible that animals not only see objects, but see them *as* having certain features (Burge 2010, pp. 34-42), and if explanation of this aspect of animal perception required the presence of a representational component, then this would seem to tell against the above suggestion. But this aspect of animal perception can be captured by the relational component of perceptual consciousness, without any appeal to a representational element, since the relata of the primitive relations of sensory apprehension which constitute perceptual consciousness include not only empirical objects but also their properties. An animal which apprehends the leaf and its particular colour sees the leaf *as* green. So the representational element to human perceptual consciousness isn't required to explain the possibility of perceiving as. An interesting implication concerns the possibility of properties which we can't be related to via sensory relations of apprehension but which nevertheless structure perceptual consciousness in such a way that we see things *as* having those properties. The possibility of seeing things *as* having these properties is only possible given the representational component to perceptual consciousness. This is important if one is sympathetic to the Kantian thought that we cannot intuit categorial properties, such as that of causality. For if we can see one thing as causing another, as Ducasse and Anscombe each claimed (Ducasse 1965; Anscombe 1971) but cannot intuit causal relations, then the explanation of this will need to turn on the representational element to perception. Thanks to the Editor for raising these issues.

 © Gomes 2016

thinking noted in §3 of this paper. But it also, I believe, provides a way of thinking about visual experience which is attractive in its own right: visual perception relates us immediately to empirical particulars through a non-conceptual mode of apprehension but, as beings with discursive intellects, this apprehension suffices for the presence of a content-bearing state which represents things as being a certain way. Contrary to the implications of Oxford Realism and non-conceptualist interpreters of Kant, naïve realism is compatible with the necessary involvement of the understanding in perceptual consciousness.[34]

5. Cognition

My aim in this paper has been to show that we can formulate a version of naïve realism in Kantian phrase. And this has involved showing the compatibility of naïve realism with the intermediate claim that the understanding is necessarily involved in the perceptual consciousness of rational beings. But one might think that Kant was committed to the stronger claim: that acts of the understanding are constitutive of perceptual consciousness. And nothing I have said has shown that naïve realism is compatible with that claim. This final section will consider this issue.

I noted two reasons for thinking that Kant held the understanding to be involved in perceptual consciousness: his account of the role synthesis plays in the representation of intuitions and his aims in the Transcendental Deduction of the Categories. Do these considerations motivate the strong claim that the understanding is constitutive of perceptual consciousness or only the intermediate claim that the understanding is necessarily involved in the perceptual consciousness

[34] How does this proposal compare to that offered by McDowell? The most important difference is that whereas McDowell takes the intentional aspect of perception to be basic, the view offered here takes the primitive relations of apprehension to be basic, thus securing a form of strong naïve realism of the kind endorsed by the Oxford Realists. But it also differs from McDowell's in the way in which it works out the intermediate claim. For McDowell, the understanding is constitutive of our perceptual consciousness alone but it shapes our perceptual consciousness in such a way that there is no highest common factor between rational and non-rational perceptual consciousness (McDowell 2011, pp. 20-1, pp. 54-7; Boyle 2012). The account offered here allows a highest common factor to rational and non-rational perceptual consciousness, namely the presence of primitive relations of apprehension. It can thus provide what McDowell's and Boyle's proposals lack: an answer to the question of what makes non-rational perceptual consciousness a form of perceptual engagement with the world. But see Boyle (2012, pp. 419-20) for an expression of suspicion as to whether this question relies on a mistaken assumption.

 © Gomes 2016

of rational beings? This textual question cannot be settled here. In what follows I present some preliminary considerations.

On synthesis: many have thought Kant's discussion of synthesis to commit him to the claim that the obtaining of an intuition constitutively depends on an act of synthesis, in particular the discussion of 'the synthesis of apprehension in intuition' at A98–9 and B160–1. And if all synthesis is undertaken by the understanding (B130) then the understanding is constitutively involved in perceptual consciousness, contrary to the intermediate claim. In contrast, the account of visual experience I have set out above holds not that intuitions depend on acts of synthesis but that *representations* of intuitions depend on acts of synthesis. This makes the role of synthesis compatible with the intermediate claim but only, one might think, by departing from Kant's texts.

Clinton Tolley (2013, pp. 122-4) has recently argued otherwise. According to Tolley, there is a distinction between the mere *having* of an intuition and the representation of an intuition *as* an intuition, and the passages noted above show only that acts of synthesis are required for the representation of an intuition as such. For example, when Kant claims that '[e]very intuition contains a manifold in itself, which however *would not be represented as such* if the mind did not distinguish the time in the succession of impressions on one another …' (A99, emphasis added), the claim is that synthesis is required for intuitions to be represented *as such* [*als ein solches vorgestellt*], not simply for them to be represented or presented [*vorgestellt*]. Similarly the requirement that the manifold of intuition be synthesised if it is 'contained in one representation' (A99) is compatible with it not being required for the existence of the intuition. This way of reading those passages makes the outcome of synthesis not an intuition but the *representation* of an intuition as on the model of visual experience proposed above.

Tolley's reading of these passages makes them compatible with a denial of the strong claim that the understanding is constitutive of perceptual consciousness. If it is to convince, we need to consider all the passages which have been thought to show that intuitions constitutively depend on acts of synthesis. I won't undertake that task here.[35] It is enough to note that there is a plausible reading of Kant's claims about synthesis which takes them to support the intermediate claim without supporting the strong claim. And if that is right, then the

[35] See Tolley (2013, pp. 122-4) for the full defence.

 © Gomes 2016

account of visual experience I have offered can be made compatible with Kant's text.[36]

The issue of the Transcendental Deduction is more difficult. The requirement here is to make sense of Kant's aim in the Deduction of justifying our use of certain a priori concepts. In order for Kant to accomplish this task, it is not enough for him to show that we must, of necessity, apply the categories to the objects of experience, for that is compatible with our application of the categories to such objects being mistaken. He needs to show, further, that the categories *really* must apply. Only so will Humean scepticism about our use of a priori concepts be neutralised.[37]

The worry about the intermediate interpretation is that it can only show that we must necessarily apply the categories to the objects of experience. On the model of visual experience presented above, for instance, the understanding is necessarily involved in perceptual consciousness, but not in the generation of intuitions. This leaves open the possibility that although we necessarily experience objects as falling under the categories, that which is given to us in intuition does not fall under the categories. Were this possibility to hold, visual experience would systematically mislead us as to the categorial nature of objects in the world. And absent a way to rule out this possibility, one might think that the Humean problem remains.

This is an important concern and one which arises for all those who want to deny the strong constitutive claim (Gomes 2014). A number of options are open in response. One is to deny that Kant needs anything more than the subjective necessity of applying the categories to answer Humean scepticism. A second is to agree that Kant wants more than that, but hold that his strategy in the Deduction is inconsistent with his view of perceptual consciousness (Hanna 2011). A more palatable

[36] Does Tolley's reading of these passages make them compatible with the weak reading? For that to be the case, it needs to be made plausible that the representation of an intuition as such – that representation which results from synthesis – is just a judgement that things are a certain way. But certain passages in §26, discussed by Tolley (pp. 123-4), suggest that the kind of representation in question is a form of 'empirical consciousness' or 'perception'. I take these to suggest that acts of synthesis still result in something perceptual, even if they don't result in intuitions, as on the intermediate view.

[37] See Gomes (2010) for a reading of the Deduction in this vein, and Gomes (2014) for an argument that defenders of the weak claim fail to make sense of the Deduction's anti-Humean aims. Kant's criticism of 'preformation-system of pure reason' views (B168) shows that the mere psychological necessity of applying the categories is not enough to justify our use of the categories.

 © **Gomes 2016**

alternative would be to show that the intermediate claim can suffice for showing that the categories really must apply.

This final option would require showing how there can be necessary dependence of the representational element of perceptual consciousness on the nature of the objects of perceptual consciousness without the understanding playing a constitutive role in intuition itself. Some passages might be thought to point in this direction. Tolley (2013, pp. 123-6) points out that Kant draws a distinction between intuition [*Anschauung*] and perception [*Wahrnehmung*], where the latter involves the subject being conscious of what is given in intuition *as* a unity (A119-20, B160). And some of Kant's phrasing suggests that he thinks his solution in the Deduction requires only that the understanding be necessarily involved in *perception*, not that it be constitutive of intuition (e.g., 'Consequently all synthesis, through which even *perception* itself becomes possible, stands under the categories …' (B161, emphasis added)). But there are other passages which seem to show elsewise – e.g., the footnote at B160 – and, regardless of the textual issue, we still require an explanation of how the Humean problem of justification can be answered without the understanding being constitutive of intuition.[38]

I won't pursue these issues here. They turn on wider issues about the structure of the Transcendental Deduction and I take them to

[38] What options are available? Proponents of the strong view often assume that it is only if acts of the understanding are constitutive of perceptual consciousness that whatever is given to us in perceptual consciousness will be such as to fall under the categories. The natural way to cash this out is in terms of existential dependence: if the objects of perceptual consciousness are existentially dependent on our faculty for cognising them, then the constitutive involvement of the understanding in perceptual consciousness will determine the necessary applicability of the categories to the objects of perceptual consciousness. But it is important to note that we don't need the existential dependence of the *objects* of perceptual consciousness in order to secure the anti-Humean conclusion: the existential dependence of the *categorial properties* would suffice. If it were the case that an object's instantiation of a categorial property depended on our faculty for cognizing that property, then this alone would show that the categories must apply to the objects of perceptual consciousness – even if the objects of perceptual consciousness didn't so depend. And although Kant sometimes suggests that the objects of perceptual consciousness existentially depend on our faculty for cognizing them (e.g., at A126-8), there are other passages where he emphasises that it is not the existence of the objects of perceptual consciousness which depends on our faculties for cognising them but only that which we know of them a priori (Bxviii; *Prol.* 4:293). This points towards a way in which the intermediate view can respond to Humean scepticism. For if it can be shown that an object's possessing categorial properties is constituted in part by subjects necessarily perceptually representing objects as having those properties, then the necessary involvement of the understanding in perceptual consciousness would go some way to showing the falsity of Humean scepticism. McDowell's (1985) account of colours and values is one model for such a procedure. Thanks to a referee for pushing me to say more.

 © Gomes 2016

indicate that an onus remains on defenders of the intermediate approach to show how such views are compatible with Kant's aims in the Transcendental Deduction – at least to the extent that those intermediate approaches are intended to be Kant's own. But my aim was only to offer an account of visual experience which was *Kantian*. And are there grounds for thinking that the account I have offered is at least that?

I noted at the start of this paper that it is common in contemporary ethics to distinguish between *Kant's* ethics and *Kantian* ethics (O'Neill 1989, Part III; Wood 2008, ch. 1.1). Kant's ethics is 'the theory that Kant himself put forward' whilst Kantian ethics is 'an ethical theory formulated in the basic spirit of Kant, drawing on and acknowledging a debt to what the author of the theory takes to be his insights in moral philosophy' (Wood 2008, p. 1). If the account of visual experience motivated in this paper is to constitute a Kantian approach to the philosophy of perception, it should similarly be formulated in the basic spirit of the first *Critique*, drawing on and acknowledging a debt to the insights of Kant's theoretical philosophy. Let me conclude by suggesting two ways in which the account I have offered does this.[39]

The first concerns the distinction between sensibility and the understanding. This distinction lies at the heart of the first *Critique*: it forms the basis for Kant's account of cognition and is central to his explanation of the source and limits of metaphysical knowledge. Kant took his predecessors to have misconstrued this distinction – he mentions Locke and Leibniz by name (A271/B327) – either by intellectualising appearances or by sensualizing the concepts of the understanding. In both cases, a difference of kind is misconstrued as a difference of degree. Kant's insight is that the two faculties cannot exchange their functions and that both are drawn upon in our cognizing the world. This forms the bedrock to Kant's theoretical philosophy.

The account of visual experience I have offered respects both parts of Kant's insight: it recognises a fundamental distinction between our sensory capacity for perceptual apprehension and our intellectual capacity for conceptual representation and explains how both are combined in our perceptual engagement with the world. Moreover it does so without sacrificing the distinctness of one or other faculty: the difference between sensible intuition and intellectual representation

[39] A third consideration, which I won't focus on, has it that visual experiences need to have a conceptual element if they are to justify beliefs. This line of thought has been pushed most forcefully by John McDowell (1994, 1998) and he sometimes ascribes it to Kant. The account offered here accords with this consideration.

 © Gomes 2016

remains one of kind and not degree. The unification of sensibility and the understanding in perceptual consciousness allows the sensory and intellectual to be combined in perception without assimilating one to the other.

The second point of contact concerns the relation between consciousness and self-consciousness. Kant is often read as endorsing a close connection between these phenomena: for at least some aspects of our mental life, being conscious of something involves being aware that one is so conscious (A98; B131f; A350; A361-7; *An* 7:127-8; ML_2 28:584). The account of visual experience offered here provides one way of making good this thought in the case of perceptual experience: visual perception involves both the relational apprehension of empirical particulars and the intentional representation of oneself apprehending the empirical particular. Consciousness of the world in perception thus involves *self-conscious* awareness of oneself consciously perceiving the world: perceptual consciousness is a form of self-consciousness (Rödl 2007; McDowell 2011).[40]

This illustrates the divergence of this Kantian model from that offered by Price. Price takes the representational element of perception to represent how things are in the world: its content concerns solely the empirical particular and its perceived properties. In contrast, the account I am offering takes the representational element to have a distinctive first-personal content: it seems to the subject as if *she* is intuiting the empirical particular. This first-personal content to visual experience arises whenever discursive subjects stand in intuitive relations to empirical particulars: perceiving an empirical object ipso facto puts one into a state with first-personal content. The result is a

[40] I have said that this way of thinking about visual experience is attractive, but one might object that the implication that perceptual consciousness is a form of self-consciousness is shown to be false by reflection on cases of immersed activity, such as that of the tennis player who reacts instinctively to the movements of his opponent. Cases of this sort have featured in the debate between Hubert Dreyfus and John McDowell over whether perceptual consciousness has an essentially first-personal character (Dreyfus 2013, pp. 28-30; cf. McDowell 2007, p. 346). McDowell defends the claim (2013b, pp. 45-6), arguing that cases of immersed activity involve a subject knowing that *she* is acting in a particular way, and although a subject cannot express such knowledge without disrupting her immersed activity, the knowledge is already present to be expressed as evidenced by her capacity to give a rational explanation of her actions. Dan Zahavi and Uriah Kriegel take a similar line in their recent defence of the claim that all conscious experience involves self-consciousness (Zahavi and Kriegel 2015, §4). If this line of response can be made plausible, then the account of visual experience offered above has resources on which to draw in accounting for cases of immersed activity. See the essays collected in Schear (2013) for further discussion of this issue. Thanks to the Editor for raising this question.

 © Gomes 2016

distinctively Kantian spin on visual experience: consciousness of the world in intuition involves self-consciousness of oneself intuiting the world. This provides a second way in which this model of visual experience is formulated in the spirit of the first *Critique*.

Let me conclude. My concern in this essay has been the reconciliation of two claims. First, that visual perception essentially involves a relation to empirical particulars. Second, that the understanding is necessarily involved in perceptual consciousness. English-speaking philosophers of perception at the start of the twentieth century took these claims to be in opposition and correspondingly formulated their naïve realist account of perception in anti-Kantian terms. My aim has been to show that this was a mistake: naïve realism can be situated within a Kantian framework which marks a distinction between sensibility and the understanding and takes both to be involved in perceiving the world. The result is naïve realism in Kantian phrase.[41]

References

Works by Kant:

References to the Critique of Pure Reason follow the standard A/B pagination and use the edition edited and translated by P. Guyer and A. Wood. Cambridge: Cambridge University Press, 1998.

All other references are to the volume and page of the Akademie edition of Kant's works, *Kants Gesammelte Schriften*, ed. Königlich Preussische Akademie der Wissenschaften, vols. 1–29 (Berlin: de Gruyter, 1902–). The titles are abbreviated as follows:

An Anthropology from a Pragmatic Point of View. In *Anthropology, History and Education*, ed. Loude R.B. and G. Zöller. Cambridge: Cambridge University Press, 2007.

BL Blomberg Logic. In *Lectures on Logic*, ed. and trans. Young J. M. Cambridge: Cambridge University Press, 1992.

[41] Many people have helped me with this paper, including Ian Phillips, Rory Madden, Nick Jones, Lee Walters, Rob Watt, Oliver Rashbrook-Cooper, Thomas Land, John Callanan, Erasmus Mayr, Max Edwards, Bill Brewer, Andrew Roche and audiences in Cambridge, Oxford, Luxembourg and York. Special thanks are owed to Craig French and Andrew Stephenson for extensive comments and discussion. Colin McLear's comments were incredibly helpful in revising the paper for publication, as were those of two anonymous referees and the then-Editor of *MIND*, Thomas Baldwin.

 © Gomes 2016

FS The False Subtlety of the Four Syllogistic Figures. In *Theoretical Philosophy 1775–1770*, ed. and trans. Walford D. and R. Meerbote. Cambridge: Cambridge University Press, 1992.

JL Jäsche Logic. In *Lectures on Logic*, ed. and trans. Young J. M. Cambridge: Cambridge University Press, 1992.

ML2 Metaphsik L_2. In *Lectures on Metaphysics*, ed. and trans. Ameriks K. and S. Naragon .Cambridge: Cambridge University Press, 1997.

OPA The Only Possible Argument in Support of a Demonstration of the Existence of God. In *Theoretical Philosophy 1775–1770*, ed. and trans. Walford D. and R. Meerbote. Cambridge: Cambridge University Press, 1992.

Prol. Prolegomena to Any Future Metaphysics. In *Theoretical Philosophy after 1781*, ed. Allison H. and P. Heath. Cambridge: Cambridge University Press, 2002.

VL Vienna Logic. In *Lectures on Logic*, ed. and trans. Young J. M. Cambridge: Cambridge University Press, 1992.

Others of Kant's texts referred to:

Kant, I. 1982. *Critique of Pure Reason: Concise Text*, trans. W. Schwarz. Aalen: Scientia Verlag Aalen.

——1996. *Critique of Pure Reason*, trans. W.S. Pluhar. Indianna: Hackett.

——1998. *Critique of Pure Reason*, ed. and trans. P. Guyer and A. Wood. Cambridge: Cambridge University Press.

Works by other authors:

Abela, P. 2002. *Kant's Empirical Realism*. Oxford: Oxford University Press.

Allais, L. 2007. Kant's idealism and the secondary quality analogy. *Journal of the History of Philosophy* 45: pp. 459–84.

——2009. Kant, non-conceptual content and the representation of space. *Journal of the History of Philosophy* 47: pp. 383–413.

——2010. Kant's argument for transcendental idealism in the transcendental aesthetic. *Proceedings of the Aristotelian Society* 110: pp. 47–75.

——2011. Idealism enough: response to Roche. *Kantian Review* 16: pp. 375–98.

 © Gomes 2016

——2012. Transcendental idealism and the transcendental deduction. In D. Schulting (ed.) *Kant's Idealism: New Interpretations of a Controversial Doctrine*. Springer.

Anscombe, G.E.M. 1965. The Intentionality of Sensation: a Grammatical Feature. In R.J. Butler (ed.) *Analytic Philosophy*. Oxford: Blackwell.

——1971. Causality and Determination. Reprinted in E. Sosa & M. Tooley (eds.) 1993. *Causation*. Oxford: Oxford University Press.

Audi, P. 2012. Grounding: Toward a Theory of the In-Virtue-Of Relation. *Journal of Philosophy* 109: pp. 685–711.

Bird, G. 1962. *Kant's Theory of Knowledge*. London: Routledge & Kegan Paul.

Boyle, M. 2012. Essentially rational animals. In G. Abel and J. Conant (eds.) *Rethinking Epistemology*. Berlin: de Gruyter.

Brewer, B. 1999. *Perception and Reason*. Oxford: Oxford University Press.

——2006. Perception and content. *European Journal of Philosophy*, 14: pp. 165–81.

——2007. Perception and its objects. *Philosophical Studies* 132: pp. 87–97.

——2011. *Perception and its Objects*. Oxford: Oxford University Press.

Broad, C.D. 1923. *Scientific Thought*. London: Kegan Paul.

——1978. *Kant: An Introduction*. Cambridge: Cambridge University Press.

Burge, T. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.

Campbell, J. 2002a. Berkeley's puzzle. In T. Gendler & J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press.

——2002b. *Reference and Consciousness*. Oxford: Oxford University Press.

Child, T.W. 1994. *Causality, Interpretation, and the Mind*. Oxford: Clarendon Press.

Cook Wilson, J. 1926. *Statement and Inference, with Other Philosophical Papers*. Oxford: Clarendon Press.

Correia F. & B. Schnieder, 2012. *Metaphysical Grounding*. Cambridge: Cambridge University Press.

 © Gomes 2016

Crane, T. 1992. The nonconceptual content of experience. In T. Crane (ed.), *The Contents of Experience.* Cambridge University Press.
——2005a. What is the problem of perception? *Synthesis Philosophica* 2: pp. 237–64.
——2005b. The problem of perception. *Stanford Encyclopedia of Philosophy.* <http://plato.stanford.edu/archives/spr2011/entries/perception-problem/>
Dancy, J. 1981. On Moral Properties. *Mind* 90: pp. 367–85.
Dickerson, A.B. 2003. *Kant on Representation and Objectivity.* Cambridge: Cambridge University Press.
Dreyfus, H.L. 2013. The Myth of the Pervasiveness of the Mental. In J. Schear (ed.) *Mind, Reason, and Being-in-the-World: The McDowell-Dreyfus Debate.* Oxford: Routledge.
Ducasse, C.J. 1965. Causation: Perceivable? Or only inferred? *Philosophy and Phenomenological Research* 26: pp. 173–9.
Fine, K. 2001. The question of realism. *Philosophers' Imprint* 1: pp. 1–30.
——2012. Guide to ground. In F. Correia & B. Schnieder (eds.), *Metaphysical Grounding.* Cambridge: Cambridge University Press.
Fish, W. 2009. *Perception, Hallucination, and Illusion.* Oxford: Oxford University Press.
Gardner, S. 1999. *Routledge Philosophy Guidebook to Kant and the Critique of Pure Reason.* London Routledge.
George, R. 1981. Kant's sensationism. *Synthese* 47: pp. 229–55.
Gomes, A. 2010. Is Kant's Transcendental Deduction of the Categories fit for purpose? *Kantian Review.* 15: pp. 118–37.
——2013. Kant and the Explanatory Role of Experience. *Kant-Studien.* 104: pp. 277–300.
——2014. Kant on Perception: Naïve Realism, Non-conceptualism and the B-Deduction. *Philosophical Quarterly.* 64: pp. 1–19.
Gomes, A. & A. Stephenson, forthcoming. On the Relation of Intuition to Cognition. In D. Schulting (ed.) *Kantian Nonconceptualism.* London: Palgrave Macmillan.
Griffith, A.W. 2012. Perception and the Categories: A Conceptualist Reading of Kant's Critique of Pure Reason. *European Journal of Philosophy* 20: pp. 193–222.
Hanna, R. 2001. *Kant and the Foundations of Analytic Philosophy.* Oxford: Oxford University Press.
——2005. Kant and nonconceptual content. *European Journal of Philosophy* 13: pp. 247–90.

 © Gomes 2016

——2006. *Kant, Science and Human Nature.* Oxford: Oxford University Press.
——2008. Kantian non-conceptualism. *Philosophical Studies* 137: pp. 41–64.
——2011. Kant's Non-Conceptualism, Rogue Objects, and The Gap in the B Deduction. *International Journal of Philosophical Studies* 19: pp. 399–415.
——2013. The Togetherness Principle, Kant's Conceptualism, and Kant's Non-Conceptualism. *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/kant-judgment/supplement1.html
Hatfield, G. 2013. Perception and sense-data. In M. Beaney (ed.) *The Oxford Handbook of The History of Analytic Philosophy.* Oxford: Oxford University Press.
Kalderon, M.E. & C. Travis, 2013. Oxford Realism. In M. Beaney (ed.) *The Oxford Handbook of The History of Analytic Philosophy.* Oxford: Oxford University Press.
Kalderon, M.E. 2011. Before the law. *Philosophical Issues* 21: pp. 219–44.
Land, T. 2013. Intuition and Judgment: How Not to Think about the Singularity of Intuition (and the Generality of Concepts) in Kant. In C. La Rocca (ed.), *Kant and Philosophy in a Cosmopolitan Sense.* Berlin: de Gruyter.
Langton, R. 1998. *Kantian Humility: Our Ignorance of Things in Themselves.* Oxford: Oxford University Press.
Longuenesse, B. 1998. *Kant and the Capacity to Judge.* Oxford: Princeton University Press.
Marion, M. 2000. Oxford Realism: Knowledge and Perception I. *British Journal for the History of Philosophy* 8: 299–338.
Martin, M.G.F. 2002a. The transparency of experience. *Mind and Language* 4: pp. 376–425.
——2002b. Particular thoughts and singular thought. In A. O'Hear (ed.), *Logic, Thought, and Language.* Cambridge: Cambridge University Press.
——2003. Sensible appearances. In T. Baldwin (ed.), *The Cambridge History of Philosophy.* Cambridge: Cambridge University Press.
——2004. The limits of self-awareness. *Philosophical Studies* 120: pp. 37–89.
McDowell, J. 1985. Values and secondary qualities. In Ted Honderich (ed.) *Morality and Objectivity.* Oxford: Routledge.
——1994. *Mind and World.* Cambridge: Harvard University Press.

 © Gomes 2016

——1998. Having the world in view: Sellars, Kant, and intentionality. *Journal of Philosophy* 95: pp. 431–92.
——2007. What Myth? *Inquiry* 50: pp. 338–51.
——2008a. Avoiding the Myth of the Given. In J. Lindgaard (ed.) *John McDowell: Experience, Norm, and Nature*. Oxford: Blackwell. Reprinted in his *Having the World in View: Essays on Kant, Hegel and Sellars*. London: Harvard University Press.
——2008b. Responses. In J. Lindgaard (ed.) *John McDowell: Experience, Norm, and Nature*. Oxford: Blackwell.
——2011. *Perception as a Capacity for Knowledge*. Marquette University Press.
——2013a. Perceptual experience: both relational and contentful. *European Journal of Philosophy* 21: pp. 144–57.
——2013b. The Myth of the Mind as Detached. In J. Schear (ed.) *Mind, Reason, and Being-in-the-World: The McDowell-Dreyfus Debate*. Oxford: Routledge.
McLear, C. 2011. Kant on animal consciousness. *Philosophers' Imprint* 11: pp. 1–16.
——2014. Comments on Stefanie Grüne's 'Blinde Anschauung'. *Virtual Critique*. <http://virtualcritique.wordpress.com/2014/08/19/colin-mclear-on-stefanie-grunes-blinde-anschauung/>
——2016. Kant on Perceptual Content. *Mind* 125: pp. 95–144.
——forthcoming. The Kantian (Non)-Conceptualism Debate. *Philosophy Compass*.
Moore, G.E. 1903a. Kant's idealism. *Proceedings of the Aristotelian Society* 4: pp. 127–40.
——1903b. The refutation of idealism. *Mind* 12: pp. 433–53.
——1913. The status of sense-data. *Proceedings of the Aristotelian Society* 14: pp. 355–80
——2011. *G. E. Moore: Early Philosophical Writings*, ed. T. Baldwin & C. Preti. Cambridge: Cambridge University Press.
Nudds, M. 2009. Recent work in perception: naïve realism and its opponents. *Analysis* 69: pp. 334–46.
O'Neill, O. 1989. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge: Cambridge University Press.
Peacocke, C. 1983. *Sense and Content: Experience, Thought, and Their Relations*. Oxford: Oxford University Press.
Phillips, I. 2005. *Experience and Intentional Content*. BPhil Thesis. University of Oxford.
Price, H.H. 1932. *Perception*. London: Methuen and Co.

 © Gomes 2016

——1947. Harold Arthur Prichard. 1871–1947. *Proceedings of the British Academy* 33: pp. 331–50.
Prichard, H.A. 1909. *Kant's Theory of Knowledge.* Oxford: Clarendon Press.
Rödl, S. 2007. *Self-Consciousness.* London: Harvard University Press.
Rosen, G. 2010. Metaphysical Dependence: Grounding and Reduction. In B. Hale & A. Hoffmann (eds.), *Modality: Metaphysics, Logic, and Epistemology.* Oxford: Oxford University Press.
Russell, B. 1910. Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society* 11: pp. 108–28.
——1912. *The Problems of Philosophy.* Oxford: Oxford University Press.
Schear, J. (ed.) 2013. *Mind, Reason, and Being-in-the-World: The McDowell-Dreyfus Debate.* Oxford: Routledge.
Schellenberg, S. 2011. Perceptual Content Defended. *Noûs* 45: pp. 714–50.
Searle, J.R. 1983. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.
Sellars, W. 1956. Empiricism and the Philosophy of Mind. Reprinted in *his Science, Perception and Reality.* London: Routledge & Kegan Paul, 1963.
——1963. Phenomenalism. In *his Science, Perception and Reality.* London: Routledge & Kegan Paul, 1963.
——1968. *Science and Metaphysics: Variations on Kantian Themes.* London: Routledge & Kegan Paul.
——1978. The Role of Imagination in Kant's Theory of Experience. Reprinted in J. Sicha (ed.) *Kant's Transcendental Metaphysics.* Atascadero CA: Ridgeview Publishing Company, 2002.
Siegel, S. 2010. Do visual experiences have contents? In B. Nanay (ed.), *Perceiving the World.* Oxford: Oxford University Press.
Soteriou, M. 2005. The subjective view of experience and its objective commitments. *Proceedings of the Aristotelian Society* 105: pp. 177–90.
——2007. Content and the stream of consciousness. *Philosophical Perspectives* 21: pp. 543–68.
——2011. Perceiving events. *Philosophical Explorations* 13: pp. 223–41.
——2013. *The Mind's Construction: The Ontology of Mind and Mental Action.* Oxford: Oxford University Press.
Stephenson, A. 2015. Kant on the Object-Dependence of Intuition and Hallucination. *The Philosophical Quarterly.* 65: pp. 486–508.

 © Gomes 2016

Steward, H. 1997. *The Ontology of Mind: Events, Processes, and States.* Oxford: Clarendon Press.

Strawson, P.F. 1966. *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason.* London: Methuen & Co.

——1979. Perception and its objects. In G.F. Macdonald (ed.) *Perception and Identity: Essays Presented to A. J. Ayer.* London: Macmillan. Reprinted in his *Philosophical Writings.* Oxford: Oxford University Press.

——1992. *Analysis and Metaphysics: An Introduction to Philosophy.* Oxford: Oxford University Press.

Thompson, M. 1972. Singular Terms and Intuitions in Kant's Epistemology. *Review of Metaphysics* 26: pp. 314–43.

Tolley, C. 2013. The Non-Conceptuality of the Content of Intuitions: A New Approach. *Kantian Review* 18: pp. 107–36.

Travis, C. 2013a. *Perception: Essays after Frege.* Oxford: Oxford University Press.

——2013b. Unlocking the Outer World. In *his Perception: Essays after Frege.* Oxford: Oxford University Press.

Tye, M. 2009. Representationalist theories of consciousness. In B. McLaughlin & A. Beckermann (eds.), *The Oxford Handbook of Philosophy of Mind.* Oxford: Oxford University Press.

Walsh, W. H. 1981. Kant's Critique of Pure Reason: Commentators in English, 1875-1945. *Journal of the History of Ideas 424:* pp. 723–37.

Westphal, K.R. 2004. *Kant's Transcendental Proof of Realism.* Cambridge: Cambridge University Press.

Wood, A.W. 2008. *Kantian Ethics.* Cambridge: Cambridge University Press.

Zahavi, D. & U. Kriegel, 2015. For-me-ness: What it is and what it is not. In ,D Dahlstrom & ,W. Hopp (eds.) *Philosophy of Mind and Phenomenology: Conceptual and Empirical Approaches.* Oxford: Routledge.

© Gomes 2016

Preservation, Commutativity and *Modus Ponens*: Two Recent Triviality Results

Jake Chandler
La Trobe University
jake.chandler@cantab.net

In a recent pair of publications, Richard Bradley has offered two novel no-go theorems involving the principle of Preservation for conditionals, which guarantees that one's prior conditional beliefs will exhibit a certain degree of inertia in the face of a change in one's non-conditional beliefs. We first note that Bradley's original discussions of these results—in which he finds motivation for rejecting Preservation, first in a principle of Commutativity, then in a doxastic analogue of the rule of *modus ponens*—are problematic in a significant number of respects. We then turn to a recent U-turn on his part, in which he winds up rescinding his commitment to *modus ponens*, on the grounds of a tension with the rule of Import-Export for conditionals. Here we offer an important positive contribution to the literature, settling the following crucial question that Bradley leaves unanswered: assuming that one gives up on full-blown *modus ponens* on the grounds of its incompatibility with Import-Export, what weakened version of the principle should one be settling for instead? Our discussion of the issue turns out to unearth an interesting connection between epistemic undermining and the apparent failures of *modus ponens* in McGee's famous counterexamples.

1. Triviality and the Ramsey test

For over two and a half decades, the so-called Ramsey test, which cashes out an agent's acceptance of a conditional of the form 'If *A* then *B*' in terms of her commitment to coming to believe that *B* upon revising her beliefs so as to accommodate *A*, has been at the centre of a lively debate regarding the appropriate formulation of the rules of rational belief dynamics.

A large proportion of this debate has been carried out against the backdrop of the AGM theory of belief change.[1] In this framework, an agent's state of mind is represented by a so-called belief set, a

[1] There is, of course, also an extensive discussion of a closely related issue in the Bayesian literature, initiated by the work of Adams (1975), followed by that of Lewis (1976) and many others. See Hájek and Hall (1994) for an excellent summary.

Mind, Vol. 126 . 502 . April 2017
© Chandler 2016
doi:10.1093/mind/fzv154 Advance Access publication 28 July 2016

consistent, deductively closed set of sentences, K, drawn from a conditional propositional language L.[2] '$K * A$' is used to denote the posterior belief set resulting from the revision of the prior belief set K by A: the rational adjustment of K to accommodate the acquisition of the belief that A. In this context, the Ramsey test can thus be succinctly formulated as follows:

(RAM) $A > B \in K$ iff $B \in K * A$[3]

where $>$ is a conditional connective, standardly intended to correspond to the natural language indicative conditional.

A number of rationality constraints on the revision function have been offered, broadly guided by the idea that the adjustment of the prior belief set ought to be as conservative as possible. Among these, the so-called AGM postulates (see Alchourón et al. 1985) provide fairly uncontroversial restrictions on the 'factual'—i.e. non-conditional—composition of the posterior belief set.

In the mid-1980s, the publication of a now famous impossibility result due to Gärdenfors (1986), brought to general attention the difficulties that the Ramsey test gives rise to when the scope of the AGM postulates is extended so as to constrain not only factual but also *conditional* posterior beliefs. A number of variants of Gärdenfors's theorem can now be found in the literature. All uncover a tension between RAM, the unrestricted versions of a small subset of the AGM postulates and a fairly strong monotonicity condition on the consequence relation in L. In what follows, we shall sketch out a particularly perspicuous version, first presented in Rott (1989).

The AGM postulates involved include first of all the so-called Inclusion condition. Inclusion states that the posterior beliefs resulting from a revision do not go beyond the joint logical consequences of the prior belief set and the input to the revision:

(INC) If $B \in K * A$, then $B \in \mathrm{Cn}(K \cup \{A\})$

[2] In the literature, K is perhaps now more standardly known as a belief 'state' and represented by an upper-case Greek letter (e.g. ψ). The term belief 'set' and the letter K are typically reserved for its factual subset.

[3] As an anonymous referee points out, it has been noted that on the present interpretation of the revision operator, $*$, RAM yields apparently counterintuitive results in the context of agents capable of autodoxastic reasoning (see Hájek and Chalmers 2007 for a presentation of the difficulties, and Leitgeb 2011 for a reply). We shall however bracket this difficulty, since it is tangential to the particular issues at stake here.

 © Chandler 2016

where Cn is a consequence function, returning the set of consequences of any given subset Γ of L, and assumed to both be supraclassical—i.e. return at least the set of classical consequences of Γ—and satisfy the deduction theorem—i.e. be such that if $B \in \mathrm{Cn}(\Gamma \cup \{A\})$, then $A \supset B$ is in Γ.

Given closure of belief sets under logical consequence, the second constraint implicated, the Vacuity postulate, states that if the input to the revision is consistent with the prior belief state, then the joint logical consequences of the prior belief set and the input to the revision do not go beyond the posterior belief set:

(VAC) If $\neg A \notin K$ and $B \in \mathrm{Cn}(K \cup \{A\})$, then $B \in K * A$

The third constraint, the Success postulate, simply guarantees that any consistent input to revision is a member of the posterior belief set:

(SUC) If $\neg A \notin \mathrm{Cn}(\varnothing)$, then $A \in K * A$[4]

Finally, we have the aforementioned monotonicity condition, which ensures that the consequences of a given set of sentences are preserved when a member of the latter is replaced by a classically logically stronger counterpart:

(MON) If $B \in \mathrm{Cn}(\Gamma \cup \{A \vee C\})$, then $B \in \mathrm{Cn}(\Gamma \cup \{A\})$

It should be noted that all these principles can be weakened by restricting the membership of B to various subsets of L. We shall denote the conditional-free fragment of L by L_0. L_{i+1} is then defined as the smallest language that includes all sentences in L_i alongside those conditionals of the form '$A > B$', such that $A \in L_0$ and $B \in L_i$. Thus, in addition to factual sentences, L_1 also includes 'flat' conditionals, of the form '$A > B$', where A and B are both factual, and L_2 includes both flat and simple right-nested conditionals, of the form '$A > B$', where A is factual and B is a flat conditional. In what follows, we shall distinguish the various corresponding weakenings of the above principles, obtained by restricting B's membership to L_i, by means of the subscript i. Whilst the use of subscripting may, at first glance, strike the reader as somewhat cumbersome, the distinctions that it

[4] INC and VAC are also commonly referred to as K*3 and K*4, respectively. SUC corresponds to a slight weakening of what is known as K*2. The latter guarantees that A is in $K * A$ even in cases in which A is inconsistent. We assume here, however, that rational belief sets are consistent.

 © Chandler 2016

enables us to draw will prove to be of utmost importance. Indeed, many principles that are perfectly innocuous in their most restricted form turn out to be highly problematic when generalized.

The general proof strategy involved in the impossibility results is the following. *Step 1*: Make a 'non-triviality' assumption about the set $\mathbb{K}$ of rationally permissible belief states, an assumption stating the existence of a particular kind of prior belief set K in $\mathbb{K}$. In the particular case of Rott's result, we have:

(NT) There exists a $K \in \mathbb{K}$ and three pairwise inconsistent sentences $A, B, C \in L_0$ such that $\neg A, \neg B, \neg C \notin K$[5]

Step 2: From this assumption, derive the result that there exist two particular ways of revising K, such that each of the resulting revised belief sets contains one of two conditionals with identical antecedents but jointly inconsistent consequents. Here it is established that $\neg A > B \in K * A \vee B$ and $\neg A > C \in K * A \vee C$. *Step 3*: Show that, as a result of this, there exists a single revised belief set—in the present instance, $K * A$–that contains both of these conditionals. The existence of such a set, however, is prohibited by RAM, which entails the following principle of Conditional Contradiction, given closure and consistency of belief sets:

(CC) If $\neg C \in \mathrm{Cn}(B)$ and $A > B \in K$, then $A > C \notin K$

The derivation in Step 2 requires very little by way of assumptions: the right-to-left half of RAM_0, SUC, INC_0 and VAC_0.[6] None of these principles seem unduly strong.[7] Step 3 is considerably more controversial. It makes use of the far more substantial INC_1, VAC_1 and MON_1 to derive a restricted version of condition known as 'Monotonicity of Additions', guaranteeing that, in the circumstances, all L_1 members of

[5] Note that, since $\neg A, \neg B, \neg C \notin K$ and the sentences A, B and C are pairwise inconsistent, we also have $A, B, C \notin K$, by virtue of closure and consistency of belief sets.

[6] See Appendix, Observation 1.

[7] It should be noted that a number of people have considered the option of weakening RAM_0. Gärdenfors (1987) shows that trouble still ensues from a number of these proposals, including the ones proposed by Rott (1986). Lindström and Rabinowicz (1992) do offer a weakening that escapes Gärdenfors's later result, but it is one that imposes extremely stringent conditions on beliefs about conditionals. Rabinowicz (1996) has also cast aspersions on VAC_0, although, for reasons that space considerations preclude us from outlining here, we do not find his discussion entirely convincing.

 © Chandler 2016

$K * A \vee B$ or $K * A \vee C$, and in particular $A > B$ and $A > C$, respectively, will find their way into $K * A$:

(MA$_1$) If $\neg A \notin K$ and $C \in K * A \vee B$, then $C \in K * A$, where $A, B \in L_0$ and $C \in L_1$[8]

It is on MA$_1$ that the majority of contributors to the literature have been inclined to pin the blame, with Lindström and Rabinowicz (1995, p. 153) offering a convincing counterexample to the principle. Beyond this, it has remained somewhat unclear, in turn, which of INC$_1$, VAC$_1$ and MON$_1$ ought to be weakened and in what way.

In a noteworthy addition to the literature, however, Etlin (2009) has shown that if one grants the additional plausible principle according to which $\neg A > B$ and $\neg A > C$ are jointly inconsistent, then a retreat from VAC$_1$ to VAC$_0$ fails to get us out of trouble: either MON$_1$ or INC$_1$ must be relinquished. Indeed, he notes that the detour via MA$_1$ can be dispensed with: $\neg A > B$ and $\neg A > C$ needn't even make it into $K * A$ to create difficulties. On the aforementioned assumption that these conditionals are jointly inconsistent, trouble would ensue from the mere fact of their both winding up in $\mathrm{Cn}(K \cup \{A\})$.[9] And to derive this result from the fact that $\neg A > B \in K * A \vee B$ and $\neg A > C \in K * A \vee C$, we simply need INC$_1$ (to first get the conditionals into $\mathrm{Cn}(K \cup \{A \vee B\})$ and $\mathrm{Cn}(K \cup \{A \vee C\})$, respectively) and MON$_1$ (to then get them both into $\mathrm{Cn}(K \cup \{A\})$).

Whilst Etlin's observation clearly alleviates the pressure placed on VAC by the Gärdenfors–Rott theorems, Bradley (2007, 2012) has recently noted, in a well-received pair of papers, that the principle is implicated in two further impossibility results. And these, he claims, *do* show that the principle is untenable in its unrestricted form.

Bradley's immediate target is a slight weakening of VAC, which goes under the name of 'Preservation':

(PRES) If $\neg A \notin K$ and $B \in K$, then $B \in K * A$

This principle equivalently states that if $B \in K$, but $B \notin K * A$, then $\neg A \in K$, that is, that one takes to be false any proposition whose truth one would take to undermine one's current beliefs.

The proofs essentially follow the three-step template outlined above, although Bradley does not assume that $\neg A > B \in K * A \vee B$

[8] See Appendix, Observation 2.

[9] Indeed, K is assumed to be consistent and such that $\neg A \notin K$. $\mathrm{Cn}(K \cup \{A\})$ is therefore itself consistent.

 © Chandler 2016

and $\neg A > C \in K * A \vee C$ *follow* from NT. Instead he makes an assumption, NT*, that there exists a $K \in \mathbb{K}$ that both has this property and satisfies the conditions spelled out in NT:

(NT*) There exists a $K \in \mathbb{K}$ and three pairwise inconsistent sentences $A, B, C \in L_0$ such that (i) $\neg A, \neg B, \neg C \notin K$ *and* (ii) both $\neg A > B \in K * A \vee B$ and $\neg A > C \in K * A \vee C$

This enables him to do away with INC_0, which, for reasons that we shall not discuss here (see Bradley 2012, pp. 153–4), he deems objectionable.[10]

In order to derive the result that $\neg A > B$ and $\neg A > C$ are in $K * A$, neither INC_1 nor MON_1 are appealed to. This important feature of the proofs is what allows Bradley to apply renewed pressure on VAC, via Preservation. In the first result, he makes use of the relatively modest $PRES_1$ and the following restricted version of a principle of Commutativity which states that, under certain conditions, the order in which a pair of successive revisions takes place does not make a difference to the contents of the resulting belief set:

(COM_1) If $\neg A, \neg C \notin K$, $\neg C \notin K * A$, and $\neg A \notin K * C$, then $B \in (K * A) * C$ iff $B \in K * A \wedge C$, where $B \in L_1$

In the second result, the work is done by the stronger $PRES_2$, alongside the following principle of closure of belief sets under *modus ponens*:

(MP_1) If $A, A > B \in K$, then $B \in K$, where $A \in L_0$ and $B \in L_1$

In both cases, Bradley initially pins the blame on Preservation, arguing for two successive weakenings of the principle.

Bradley's first result is the subject of §2. We first outline a correction to his proof, noting that, contrary to what he claims, a weak version of INC, namely INC_0, *would* in fact appear to be required to derive the relevant result. We then suggest that COM_1 is excessively strong, noting that it is inconsistent with an uncontroversial weakening of a popular principle of 'iterated' belief revision.

In §3, we then turn to Bradley's second result. As we have just mentioned, he initially intends to resolve the tension by weakening

[10] It should perhaps be noted that Bradley (2007, p. 5) also makes use of a somewhat stronger *logical* principle of Conditional Contradiction rather than a doxastic one such as the one presented above. The idea is essentially to declare as jointly *inconsistent*, rather than simply jointly unbelievable, those conditionals with identical, consistent antecedents and jointly inconsistent consequents. Something similar to our principle CC is then recovered via the consistency requirement on belief sets.

 © Chandler 2016

$PRES_2$. We point out in §3.1, however, that the principle that he winds up recommending is both too strong and too weak. In a later paper he opts for rejecting MP_1, on the grounds of its incompatibility with the rule of Import-Export. He does not, however resolve the issue of the extent to which it ought to be weakened. This is the focus of §3.2, in which we offer the key positive contribution of the present paper. We note that a slightly restricted version of MP_1, which we derive from weak principles of belief dynamics, would provide a suitable replacement. This suggestion turns out to highlight an interesting potential connection between epistemic undermining and counterexamples to *modus ponens*.

In order not to overload the main body of the article, a number of proofs have been relegated to the appendix.

2. Triviality from Commutativity

As mentioned in the previous section, Bradley offers what he takes to be a proof that, in the presence of CC and NT*, $PRES_1$ is inconsistent with COM_1:

> [L]et $\{A, B, C\}$ be a set of mutually contradictory … sentences and suppose that an agent's initial epistemic state is given by a set K containing none of $A, \neg A, B, \neg B, C$ or $\neg C$. Suppose also that $\neg A > B$ belongs to $K * A \vee B$ and that $\neg A > C$ belongs to $K * A \vee C$. Now consider the sets $(K * A \vee B) * A$ and $(K * A \vee C) * A$, representing the epistemic states resulting from learning or supposing that A when in the states represented by $K * A \vee B$ and $K * A \vee C$. Since $\neg A$ does not belong to either $K * A \vee B$ or $K * A \vee C$, it follows from $[PRES_1]$ that $\neg A > B$ belongs to $(K * A \vee B) * A$ and that $\neg A > C$ belongs to $(K * A \vee C) * A$ and from $[COM_1]$ that $[\neg A > B$ belongs to $(K * A \vee B) * A$ iff it belongs to $K * A$ and $\neg A > C$ belongs to $(K * A \vee C) * A$ iff it belongs to $K * A]$. Hence both $\neg A > B$ and $\neg A > C$ belong to $K * A$, contrary to the hypothesis that it is not permissible to believe both that $\neg A > B$ and that $\neg A > C$. (Bradley 2007, p. 12)

The upshot of this, Bradley tells us, is that, assuming the legitimacy of CC and NT*, we must give up either COM_1 or $PRES_1$. He aims to keep the first, but only retain a substantially weakened version of the second. He recommends a retreat to

($PRES_1$*) If $\neg A \notin K$ and $B > C \in K$, then $B > C \in K * A$, so long as $\neg B \notin \mathrm{Cn}(A)$

which states that conditional beliefs are preserved under non-belief-contravening revisions so long as their antecedent is consistent with

 © Chandler 2016

the revision-inducing input. Bradley remarks that, in the absence of any further background assumptions, a shift from PRES_1 to PRES_1^* allows one to steer clear of a conflict between NT* and COM_1: since $\neg\neg A \in \text{Cn}(A)$, PRES_1^* is too weak to enable the derivation of $\neg A > B \in (K * A \vee B) * A$ and $\neg A > C \in (K * A \vee C) * A$ in the proof above.

Before turning to more substantial issues, we first note a lacuna in the proof sketched out above. Indeed, the claim that $\neg A$ does not belong to either $K * A \vee B$ or $K * A \vee C$, which plays an essential role in the proof, remains unsubstantiated. And it would appear that to derive it, we require INC_o. Here is how the amended proof might go. Assume NT*. Assume for *reductio* that $\neg A \in K * A \vee B$. By INC_o, we then have $\neg A \in \text{Cn}(K \cup \{A \vee B\})$. Since Cn is assumed to satisfy the deduction theorem, it follows that $(A \vee B) \supset \neg A \in K$. Since $\neg A \in \text{Cn}((A \vee B) \supset \neg A)$, we then have, by closure of belief sets, $\neg A \in K$, in contradiction with our initial assumption. By parallel reasoning, we can establish that $\neg A \notin K * A \vee C$. QED.[11]

Let us now turn to the moral that Bradley would have us draw from the result, namely, that COM_1 ought to be retained at the expense of PRES_1.

Now COM_o, the restriction of COM to factual sentences, is both rather weak and rather plausible. It is widely endorsed in the belief revision community, and follows from a subset of the AGM and so-called DP postulates.[12] COM_1, however, is substantially stronger. It is not just inconsistent with its being *mandatory* that both $\neg A > B$ and $\neg A > C$ are preserved in the transitions from $K * A \vee B$ to $(K * A \vee B) * A$ and from $K * A \vee C$ to $(K * A \vee C) * A$, respectively, as PRES_1 entails. As Bradley's result demonstrates, it is also inconsistent with its being merely *permissible*.

This strikes us as being too restrictive. To illustrate, consider an example discussed by Bradley himself. You have in front of you three urns, one of which contains a prize. You are initially agnostic as to the location of the prize. Scenario 1: You obtain information that suggests that the prize is either in urn A or in urn B, and thus come to hold that

[11] In a second version of the proof, Bradley eschews the aforementioned claim altogether. Instead, he appeals to the observation that A does not contradict either $A \vee B$ or $A \vee C$ to derive $\neg A > B$ and $\neg A > C$'s membership of $(K * A \vee B) * A$ and $(K * A \vee C) * A$, respectively (Bradley 2012, p. 152). PRES, however, does not license this inference.

[12] See Appendix, Observation 3. The DP postulates, introduced by Darwiche and Pearl (1997) and widely endorsed in the belief revision community, extend the constraints already imposed by the AGM postulates.

 © Chandler 2016

if the prize is not in urn *A*, it is in urn *B*. Scenario 2: You obtain information that suggests that the prize is either in urn *A* or in urn *C*, and thus come to hold that if the prize is not in urn *A*, it is in urn *C*. Now, in either scenario, upon obtaining further information suggesting that the prize is in urn *A*, it would not seem intrinsically irrational for you to preserve your commitment to the relevant conditional. In Scenario 1, for instance, coming to believe that the prize is in urn *A* seems rationally compatible with maintaining that if it turns out that the prize is not in *A* after all, it is in urn *B*. COM_1, however, says otherwise: the conditional would have to be dropped in at least one of the two situations.

Furthermore, it should be noted that even the comparatively strong proposition that the preservation of $\neg A > B$ and $\neg A > C$ is *required* by rationality might turn out to enjoy some support in the belief revision community. Indeed, the claim would follow from the conjunction of RAM and

(WDP_2) If C is equivalent to $\neg A$, then $B \in K * A$ iff $B \in (K * C) * A$, where $A, B, C \in L_0$. [13]

WDP2 is a substantial weakening of a principle satisfied by the vast majority of proposals regarding so-called iterated revision, namely Darwiche and Pearl's postulate DP2, which states that if $\neg C \in \mathrm{Cn}(A)$, then $B \in K * A$ iff $B \in (K * C) * A$, where $A, B, C \in L_0$.[14] Those familiar with the literature may wish to note that WDP2 remains weak enough to avoid the objection levelled at DP2 by Cantwell (1999, pp. 71–3), who asks us to consider the following case:

> You initially believe that George is a mass murderer [(*B*)] and therefore not a law-abiding citizen [($\neg C$)] but suspend judgment as to whether or whether or not he is a shoplifter [(*A*)]. You then come to believe that he is in fact a law-abiding citizen, and hence not a mass murderer. But then the news comes in that George is in fact a shoplifter, and hence not a law-abiding citizen after all.

On the assumption that you would initially have still believed that George is a mass murderer had you found out that he is a shoplifter ($B \in K * A$), since $\neg C \in \mathrm{Cn}(A)$, DP2 entails that you ultimately

[13] If we assume the plausible principle of Extensionality (EXT), according to which respective revisions by two logically equivalent sentences yield the same posterior belief set, WDP2 can be more simply recast as: $B \in K * A$ iff $B \in (K * \neg A) * A$.

[14] DP2 is notably satisfied by lexicographic revision (Nayak 1994), restrained revision (Booth and Meyer 2006) and natural revision (Boutillier 1996).

 © Chandler 2016

recover your belief that George is a mass murderer ($B \in (K * C) * A$). This, however, is a result that Cantwell deems counterintuitive: DP2, according to him, must go. WDP2, however, does not license the conclusion that $B \in (K * C) * A$, since the requirement that $A \in \mathrm{Cn}(\neg C)$ is not satisfied (some law-breakers have respect for the retail industry).

3. Triviality from *modus ponens*

In a second no-go theorem, Bradley effectively shows that, in the presence of SUC, CC, the right-to-left direction of RAM_1 and NT*, PRES_2 is inconsistent with MP_1 (Bradley 2007, pp. 8–9). Indeed, assume NT*. From the fact that $\neg A > B \in K * A \vee B$ and $\neg A > C \in K * A \vee C$, we recover, by the right-to-left direction of RAM_1, the result that $(A \vee B) > (\neg A > B)$ and $(A \vee C) > (\neg A > C)$ are both in K. This, in turn, in conjunction with the fact that $\neg A$ is not in K, yields, by PRES_2:

(1) $(A \vee B) > (\neg A > B) \in K * A$ and $(A \vee C) > (\neg A > C) \in K * A$

By SUC it follows that A is in $K * A$, and hence, by closure of belief sets, that

(2) $A \vee B \in K * A$ and $A \vee C \in K * A$

But it then follows from (1) and (2), by MP_1, that

(3) $\neg A > B \in K * A$ and $\neg A > C \in K * A$

in contradiction with CC. QED

3.1 Weakening Preservation

Bradley is initially inclined to see in this result a motivation for rejecting PRES_2 (Bradley 2007, p. 12) and offers a suggestion as to how the principle should be weakened so as to avoid the derivation of proposition (1).[15] The proposal is that $(A \vee B) > (\neg A > B)$ and $(A \vee C) > (\neg A > C)$ might fail to be in $K * A$ because the antecedent of the embedded conditional, $\neg A$, is inconsistent with the sentence

[15] Another option would of course be to reject the right-to-left direction of RAM_1, so as to block the derivation of $(A \vee B) > (\neg A > B), (A \vee C) > (\neg A > C) \in K$. We concur with Bradley (2007, p. 9), however, that it is not the inclusion of these nested conditionals in K that is problematic here.

 © Chandler 2016

with which K is being revised, namely, A. More precisely, the suggestion would be, where B, C, D are in L_o:

(PRES_2^*) If $B > (C > D) \in K$ then $B > (C > D) \in K * A$ so long as (i) $\neg A \notin K$ and (ii) $\neg C \notin \text{Cn}(A)$[16]

But, in retreating to PRES_2^*, Bradley appears to have both given up too much and given up too little, endorsing a principle that is both too weak and too strong.

Regarding excessive strength, consider first the following example. My brother has just joined us on holiday, where we are planning to take diving classes. Let 'A' stand for the proposition that his wristwatch looks very much like a brand new Luminox from the Navy Seal range, 'B' for the proposition that the watch is the genuine article, 'C' for the proposition that he bought the watch from a street vendor when transiting through Bangkok, and 'D' for the proposition that the watch will survive the week's nautical antics. Let K be my state of belief prior to seeing his recent horological acquisition, in which I am completely agnostic as to whether or not A, B, C, D, or any contingent Boolean combination thereof. I would take A to provide a sufficient but defeasible reason to believe B, and B, in turn, to bear the same relation to the belief that D. So B is in $K * A$ and hence $(K * A) * B$ is identical to $K * A$. But I also would take C to undercut the support

[16] This is not in fact quite what he proposes in print. What he does is to restrict Preservation in such a way that a conditional $B > C$ is not guaranteed to be preserved in the transition from K to $K * A$ if it is not the case that $C \in (K + B) * A$, where '$K + B$' denotes $\text{Cn}(K \cup \{B\})$, $B \in L_0$ and $C \in L_1$ (Bradley 2012, p. 158). In the context of his already weakened Preservation condition PRES_2^*, this yields:

> If (i) $\neg A \notin K$, (ii) $\neg B \notin \text{Cn}(A)$, and (iii) $C \in (K + B) * A$, then if $B > C \in K$, then $B > C \in K * A$

The thought is that, in relation to the triviality result, whilst conditions (i) and (ii) are met, clause (iii) is not satisfied: $\neg A > B \notin (K + A \vee B) * A$. The reasoning goes as follows. Either $\neg A > B$ is in $K + A \vee B$ or it is not. If it is not, the above principle clearly cannot yield the result that $\neg A > B$ is in $(K + A \vee B) * A$, since there is no conditional to preserve in the first place. But neither can it yield this result if it is in $K + A \vee B$, since the input to revision (A) is inconsistent with the antecedent of the conditional ($\neg A$), yielding a failure of condition (ii).

Setting aside worries that $K + B$ might fail to constitute a legitimate belief set, and thus $(K + B) * A$ fail to be defined (see Rott 1989 and Hansson 1992, also cited in Bradley 2007 p. 10), it is not particularly clear why Bradley requires C to be in $(K + B) * A$, rather than merely in $K * A$. The constraint is certainly not motivated by the triviality result, since the above reasoning carries over, *mutatis mutandis*, to the variant of the proposal in which (iii) is replaced by the requirement that $C \in K * A$.

In correspondence, Bradley has suggested that PRES_2^* might better capture his original intention.

 © Chandler 2016

that A provides for B. So, whilst D is in $K * A$, as I infer B from A and then D from B, D is not in $(K * A) * C$ and hence not in $((K * A) * B) * C$: C undercuts the probative force of the only grounds I had to hold that B, which, in turn, constituted my only reason to hold that D. Since I do not take C to undercut the support that B provides for D (a Navy Seal watch is water-resistant, whatever its place of purchase), and B remains in $(K * B) * C$, we have the result that D is in $(K * B) * C$. So, to recapitulate, D is in $(K * B) * C$, but *not* in $((K * A) * B) * C$: in terms of conditionals, given the Ramsey test, we would have $B > (C > D)$ in K but not in $K * A$. However, since, in addition, $\neg C$ is not a consequence of A and $\neg A$ is not in K, Bradley's preservation principle would judge my doxastic dispositions to be irrational.

There is also a technical result that could be taken to show that PRES_2^* is too strong. In a recent paper, Rott (2011, Corollary 6) shows that the somewhat more demanding principle PRES_2 has some very unwelcome consequences in the presence of the AGM postulates and RAM_1. Interestingly, the proof that he provides can be tweaked to show that even PRES_2^* itself sits unhappily alongside these assumptions. More specifically, it can be proven that PRES_2^* is inconsistent with NT in the presence of INC_o, VAC_o, SUC and RAM_1: hanging on to Bradley's suggested principle would plausibly require giving up on the Ramsey test for iterated conditionals.[17]

Regarding excessive weakness, PRES_2^* leaves us unable, given the Ramsey test, to predict the preservation of a number of right-nested conditionals. Suppose, for instance, that Eeyore believes that if he and Pooh decide to have a picnic the following Sunday (A), then if it proves to be a rainy day (B), then they'll be sorry (C): $A > (B > C) \in K$. This belief ought to be preserved in the event that his outlook on the weather takes an optimistic turn: $A > (B > C) \in K * \neg B$. PRES_2^* fails to predict this, since $\neg C$ is not a consequence of A.

3.2 Weakening modus ponens: MP for Import-Exporters

It should be noted, however, that Bradley appears to have now backtracked on his original diagnosis of the second triviality result. Indeed, in his most recent discussion of the result (Bradley 2012, p. 152), he expresses doubts regarding MP_1, suggesting that the scope of MP

[17] Whether or not this would be too great a price to pay is up for discussion, of course. See Appendix, Observation 4 for a derivation of the result.

 © Chandler 2016

should be restricted to cases in which $B \in L_0$. In his current view, '*Modus ponens* is invalid when applied to nested conditionals and for this reason there is good reason for rejecting the argument [from the above no-go theorem to the falsity of $PRES_2$]'.

So why the change of heart? As it happens, there exists a well-known alternative derivation of proposition (1) above. Since this line of argument does not appeal to $PRES_2$, it potentially provides independent grounds for the rejection of MP_1. Indeed, proposition (1) can be quickly established using RAM_0, alongside the following principle of doxastic Import-Export:

(IE) $A > (B > C) \in K$ iff $(A \wedge B) > C \in K$.[18]

The upshot of this is that, to the extent that IE is deemed more plausible than MP_1, the latter must be rejected.[19] Furthermore, as Bradley (2007, pp. 11–12) notes, rejecting MP_1 would mean that we would have to also jettison the left-to-right direction of RAM_1, in turn, at least to the extent that we are committed to a rather intuitive Identity principle, which states that revision by a sentence that is already believed yields no change in beliefs, factual or otherwise:

(ID) If $A \in K$, then $K * A = K$[20]

Indeed, assume that $A > B \in K$, where $B \in L_1$. By the left-to-right direction of RAM_1, we have $B \in K * A$. Assume that $A \in K$. By ID, we have $K * A = A$, and hence $B \in K$. QED.

Bradley, however, never ends up addressing a rather crucial question, namely, if one does insist on IE, exactly *how much* of MP_1 and RAM_1 should one be giving up? As it turns out, there is a fairly substantial restriction on the extent to which these principles can be weakened, if one assumes RAM_0 and IE. Indeed, given these two principles and a number of standard constraints on belief dynamics,

[18] In fact, only the right-to-left direction of this principle, the so-called rule of Export, is needed here. See Appendix, Observation 5.

[19] This rejection of MP_1 on the basis of IE is of course closely related to McGee's well-known objection to the former's logical counterpart (McGee 1985).

[20] A restricted version of this principle, namely ID_0, in which the identity is required to hold only between the L_0-subsets of $K * A$ and K, follows from the conjunction of INC_0 and VAC_0 (proof omitted, but trivial). It should be noted that the full-blown version is violated, in the presence of the Ramsey test, by a number of proposals for handling so-called iterated revision, including lexicographic revision and restrained revision. Such a violation is regrettable, however, in our view, and the aforementioned proposals can easily be modified so as to accommodate the principle.

 © Chandler 2016

it can be shown that the following weakening of the left-to-right direction of RAM_1 is recovered:

($\mathrm{WRAM}_1^{\rightarrow}$) If $A \in (K * A) * B$ or $B > C \in K * \neg A$ then, if $A > (B > C) \in K$, then $C \in (K * A) * B$

where $A, B, C \in L_0$.[21] Given ID, this in turn yields the following restricted version of MP_1:

(WMP_1) If $A \in (K * A) * B$ or $B > C \in K * \neg A$ then, if $A, A > (B > C) \in K$, then $B > C \in K$

where again $A, B, C \in L_0$. Whilst clearly stronger than MP_0, this last principle remains weak enough to avoid the derivation of (3) from the conjunction of (1) and (2) in the triviality result. For the inference to go through, one would need to establish at least one of two claims, namely, that either (a) $A \vee B \in ((K * A) * A \vee B) * \neg A$ and $A \vee C \in ((K * A) * A \vee C) * \neg A$ or (b) $\neg A > B \in K * \neg(A \vee B)$ and $\neg A > C \in K * \neg(A \vee C)$. Both (a) and (b), however, can be shown to be inconsistent with a set of extremely weak assumptions.[22]

That is not all. Indeed, one might still wonder whether WMP_1 is the strongest weakening of MP_1 that can be recovered, in the sense that the *falsity* of (a) and (b) would then suffice for the *failure* of MP_1. It turns out that this question can be answered in the positive: one can show that, given RAM_0 and a set of standard assumptions regarding belief dynamics, for all $A, B, C \in L_0$, if $A \notin (K * A) * B$ and $B > C \notin K * \neg A$, then $B > C \notin K$, even when $A, A > (B > C) \in K$.[23,24] Note in passing that one upshot of this is that, when one's belief set K is such that $A \notin (K * A) * B$, one cannot be properly said to *infer* the conclusion $B > C$ from the minor premiss A, assuming the major premiss $A > (B > C)$, since the conclusion will only be rationally believable to the extent that it would be so even if the minor premiss were disbelieved.

These results are certainly of technical interest, but it is also worth taking a closer look at the first disjunct of the antecedent of WMP_1,

[21] See Appendix, Observation 6.

[22] See Appendix, Observation 7.

[23] See Appendix, Observation 8. I am grateful to an anonymous referee for pressing me to settle this.

[24] Note that this also demonstrates that $\mathrm{WRAM}_1^{\rightarrow}$ is also the strongest weakening of the left-to-right direction of RAM_1.

 © Chandler 2016

namely, $A \in (K * A) * B$, since it admits a rather interesting intuitive informal interpretation.

In a recent article (Chandler 2013), I have argued in favour of a Ramsey-test-style principle, not for beliefs about conditionals, but for beliefs about *reasons for belief*. The proposal was that to check whether or not one is committed to the belief that A is a reason to believe B (notation: $A \triangleright B \in K$), one should first hypothetically retract one's commitment to B, then hypothetically commit oneself to A, upon which one should verify whether or not a commitment to B is recovered in the process. In other words, one ought to check whether or not $B \in (K \dot{-} B) * A$. The episode of contraction by B prior to revision by A, which distinguishes the proposal from the Ramsey test, is crucial to capturing the idea that coming to believe A should be *relevant* to acquiring the belief that B. Indeed, as is well known, under mild conditions, the use of an exact analogue of the Ramsey test for evidential beliefs would entail a 'centring' condition, according to which, for *any* A and B, no matter how intuitively unrelated, if one believes both A and B, then one thereby believes that A is a reason to believe B (Gärdenfors 1979). Taking into consideration the appropriately contracted belief set blocks the derivation of this result.

But this suggestion regarding reasons to believe suggests a very natural parallel proposal regarding beliefs about reasons *not* to believe—'underminers', as one could call them: in order to verify whether or not one is committed to the claim that A provides a reason not to believe that B (let us denote this by '$A \triangleright \mid B \in K$'), one should first hypothetically *commit* oneself to B, then hypothetically commit oneself to A, upon which one should verify whether or not one's commitment to B *disappears* in the process. In other words, one ought to check whether or not $B \notin (K * B) * A$. Again, the episode of revision by B prior to revision by A is essential to the proposal. Indeed, were one simply to propose to check whether or not $B \notin K * A$, we would, under mild conditions, recover the unhappy result that everything that one believes is taken to be a reason not to believe anything that one does not believe.[25] Again, taking into consideration the appropriately revised belief set blocks the derivation of this result, securing a requirement of relevance between A and B.[26]

[25] See Appendix, Observation 9.

[26] In fact, one can show that such a result, when coupled with our proposal, would entail under weak conditions that if $A \in K$ and $B \notin K$, then $A \notin K * B$. See Appendix, Observation 10.

 © Chandler 2016

Returning to our weakening of MP_1, we can see that it can be plausibly interpreted as stating that MP_1 holds on condition that either (a) the agent does not take the antecedent of the embedded conditional to undermine—i.e. to provide a reason not to believe—the antecedent of the whole conditional ($B \triangleright | A \notin K$), or (b) the agent *does* take such a relation of undermining to obtain ($B \triangleright | A \in K$), but would remain committed to the embedded conditional *even if* he or she were to come to believe that the antecedent of the whole conditional is false ($B > C \in K * \neg A$). Similarly, our sufficient condition for failure of MP_1 can be interpreted as stating that the latter fails when the relevant undermining relation is taken to obtain, but the embedded conditional would not be endorsed upon coming to believe the falsity of the antecedent of the whole conditional ($B > C \notin K * \neg A$). This is rather intuitive. Indeed, assume that the assessment of $B > C$ goes by checking that C is present upon revising one's beliefs by B. If one takes B to undermine A, then a revision of one's beliefs by B will leave us unable to count on the support of A when it comes to ensuring that C. For $B > C$ to be accepted, B is going to have to stand on its own two feet, so to speak, and be able to deliver C whether A is taken to be true or taken to be false.

This interpretation of whether or not $A \in (K * A) * B$ in terms of whether or not B is taken to undermine A is noteworthy, since it would offer an intuitive diagnosis of the apparent failure of *modus ponens* in McGee's widely discussed election case. McGee writes:

> Opinion polls taken just before the 1980 election showed the Republican Ronald Reagan decisively ahead of the Democrat Jimmy Carter, with the other Republican in the race, John Anderson, a distant third. Those apprised of the poll results believed, with good reason:
>
> If a Republican wins the election, then if it's not Reagan who wins it will be Anderson.
>
> A Republican will win the election.
>
> Yet they did not have reason to believe
>
> If it's not Reagan who wins, it will be Anderson. (McGee 1985, p. 462)

Clearly, since those apprised of the poll will only hold that a Republican will win the election on the basis of a belief that Reagan will win the election, it follows that they would take the news that Reagan will not win as undermining their belief in a Republican victory. Furthermore, it is plausible that were these people to come to believe that Reagan will not win the election, they would not believe

 © Chandler 2016

that if Reagan loses, then Anderson will win. So the inference to McGee's final conditional would not be licensed on the mere basis of the previous two.

Furthermore, it is rather plausible that, were such people to come to revise their beliefs so as to expect a Republican defeat, they would not believe that if Reagan will lose, then Anderson will win. If this is the case, then according to the above analysis, not only would they not have a reason to believe the final conditional, as McGee suggests, but they would in fact be rationally prohibited from doing so.

4. Concluding remarks

As we have seen, it would appear that Bradley's first result fails to have the probative force advertised. The jury therefore remains out as to whether the weaker form of Preservation for conditionals, PRES_1, is a sensible constraint on belief dynamics.

Regarding the implications of the second result, pertaining to the stronger principle PRES_2, things are perhaps less clear-cut. Much depends on the relative plausibilities of IE and MP_1. What is certainly true is that, to the extent that one wishes to endorse MP_1 in its unqualified incarnation, the proposed retreat to PRES_2^* is not a promising move. A better proposal is needed. Furthermore, for those who wish to endorse IE, belief revision theory offers clear guidance as to how MP_1 should be accordingly weakened, enabling us to derive a principle that also provides an intuitive account of the apparent failure of *modus ponens* in McGee's election case.[27]

5. Appendix

In what follows, we make occasional use of the following further principles, in addition to some of the ones already outlined in the main body of the paper:

(SUPEX) If $B \in K * (A \wedge C)$, then $B \in \text{Cn}((K * A) \cup \{C\})$

[27] I am very grateful to Igor Douven, David Etlin, Hannes Leitgeb and Greg Wheeler for their valuable comments on earlier drafts of this paper, as well as to Richard Bradley for his helpful correspondence. Three anonymous referees for this journal also provided extremely detailed feedback that has resulted in a much improved end result. As pointed out above, I owe some important steps in the derivation of Observation 6 to Richard Booth. The research underpinning this article was funded by, and carried out at, the Center for Mathematical Philosophy of LMU Munich. I would like to thank Hannes Leitgeb and Stephan Hartmann for providing such a truly exceptional and stimulating working environment.

 © Chandler 2016

(SUPEX) If $\neg C \notin K * A$ and $B \in \mathrm{Cn}((K * A) \cup \{C\})$, then $B \in K * (A \wedge C)$

(DP_1) If $C \in \mathrm{Cn}(A)$, then $B \in K * A$ iff $B \in (K * C) * A$

(DP_3) If $B \in K * A$, then $B \in (K * B) * A$

(DP_4) If $\neg B \notin K * A$, then $\neg B \notin (K * B) * A$

The first two 'subexpansion' and 'superexpansion' postulates are also individually known as 'K*7' and 'K*8' respectively, and collectively as the 'supplementary' AGM postulates. Again, we use the subscript i to denote the restriction of these principles to B's that are members of L_i. The last three conditions are non-controversial members of the Darwiche-Pearl postulates, mentioned above in footnote 12. They are only assumed to hold for $B \in L_0$.

Observation 1. *Given the right-to-left half of* RAM_0, INC_0, VAC_0 *and SUC, if* $K \in \mathbb{K}$ *is such that, for pairwise inconsistent* $A, B, C \in L$, *we have* $\neg A, \neg B, \neg C \notin K$, *then* $\neg A > B \in K * A \vee B$ *and* $\neg A > C \in K * A \vee C$

Proof of Observation 1. We show that $\neg A > B \in K * A \vee B$. The proof that $\neg A > C \in K * A \vee C$ is analogous. Given that A and B are mutually exclusive, $(A \vee B) \supset A$ is equivalent to $\neg B$. Since $\neg B \notin K$, it then follows by closure of belief sets that $(A \vee B) \supset A \notin K$. Since Cn is assumed to satisfy the deduction theorem, we therefore have $A \notin \mathrm{Cn}(K \cup \{A \vee B)\})$. Hence, by INC_0, we have (1) $A \notin K * A \vee B$. Since $A \vee B$ is consistent, by SUC, we have $A \vee B \in K * A \vee B$, and hence, by (1) and VAC_0, we have (2) $A \vee B \in (K * A \vee B) * \neg A$. Since $\neg A$ is consistent, by SUC, we have $\neg A \in (K * A \vee B) * \neg A$, and hence, by (2) and closure of belief sets, we recover (3) $B \in (K * A \vee B) * \neg A$. Finally, from (3) and the right-to-left half of RAM_0, we then obtain $A > B \in K * A \vee B$, as desired. □

Observation 2. INC_1, VAC_1 *and* MON_1 *jointly entail* MA_1

Proof of Observation 2. Consider an arbitrary $C \in L_1$ and $K \in \mathbb{K}$, such that $\neg A \notin K$ and $C \in K * A \vee B$. From the latter, by INC_1, it follows that $C \in \mathrm{Cn}(K \cup \{A \vee B\})$. By MON_1, we have $C \in \mathrm{Cn}(K \cup \{A \vee B, A\})$. If we assume that the consequence relation is supraclassical, the latter is equal to $\mathrm{Cn}(K \cup \{A\})$. Since, by hypothesis, $\neg A \notin K$, and hence, by closure of belief sets, $\neg(A \vee B) \notin K$, we then have, by VAC_1, the result that $C \in K * A$. □

© Chandler 2016

Observation 3. *VAC_o, SUC, $SUPEX_o$, $SUBEX_o$, and DP1 jointly entail COM_o*

Proof of Observation 3. Let B be a sentence in L_o, and simply assume that $\neg C \notin K * A$. By SUC, $A \in K * A$. It then follows by VAC_o that $A \in (K * A) * C$. By $SUPEX_o$ and $SUBEX_o$, it follows that $B \in (K * A) * C$ iff $B \in (K * A) * A \wedge C$. But, since $A \in Cn(A \wedge C)$, DP1 entails that $B \in (K * A) * A \wedge C$ iff $B \in K * A \wedge C$. Hence $B \in (K * A) * C$ iff $B \in K * A \wedge B$, as required. □

Observation 4. *$PRES^*_2$ and NT are jointly inconsistent in the presence of INC_o, VAC_o, SUC and RAM_1.*

Proof of Observation 4 (adapted from the proof of Corollary 6 in Rott 2011). Assume NT. We will show that, given INC_o, VAC_o, SUC and RAM_1, (i) $(A \vee B) > (\neg A > B)$ is in K, but (ii) it is not in $K * A \vee C$, in contradiction with $PRES^*_2$.

As we have already shown in Observation 1, given the right-to-left direction of RAM_o, INC_o, VAC_o and SUC, we have (1) $\neg A > B \in K * A \vee B$ and (2) $\neg A > C \in K * A \vee C$.

Regarding (i): From (1), by the right-to-left direction of RAM_1, we recover $(A \vee B) > (\neg A > B) \in K$, as desired.

Regarding (ii): Assume for *reductio* that $\neg(A \vee B) \in K * A \vee C$. By INC_o, $\neg(A \vee B) \in Cn(K \cup \{A \vee C\})$. By the deduction theorem, $(A \vee C) \supset \neg(A \vee B) \in K$. By closure of belief sets, we then have $\neg A \in K$, contrary to our assumption. So (3) $\neg(A \vee B) \notin K * A \vee C$. Since, by assumption, $\neg C \notin K$, it follows by closure of belief sets that C is consistent, and hence that (4) $\neg(B \vee C) \notin Cn(A \vee C)$. From (2), (3) and (4), it then follows by $PRES^*_2$ that $\neg A > C \in (K * A \vee C) * A \vee B$, and hence, given the left-to-right direction of RAM_o, that (5) $C \in ((K * A \vee C) * A \vee B) * \neg A$. Now, finally, assume for *reductio* that $(A \vee B) > (\neg A > B) \in K * A \vee C$. It follows by the left-to-right direction of RAM_1 that $B \in ((K * A \vee C) * A \vee B) * \neg A$. Since B and C are jointly inconsistent, by (5) and closure and consistency of belief sets, we have a contradiction. Hence $(A \vee B) > (\neg A > B) \notin K * A \vee C$, as desired. □

Observation 5. *Given RAM_o, SUC, and the right-to-left direction of IE, if, for pairwise inconsistent $A, B, C \in L_o$, we have $\neg A, \neg B, \neg C \notin K$, then $(A \vee B) > (\neg A > B) \in K * A$ and $(A \vee C) > (\neg A > C) \in K * A$*

Proof of Observation 5. Since $\{\neg A, A \vee B\}$ and $\{\neg A, A \vee C\}$ are consistent and respectively entail B and C, it follows, by RAM_o, SUC and

 © Chandler 2016

closure of belief sets, that $((A \vee B) \wedge \neg A) > B$ and $((A \vee C) \wedge \neg A) > C$ are in $K * A$, and hence, by Export, so too are $(A \vee B) > (\neg A > B)$ and $(A \vee C) > (\neg A > C)$. □

Observation 6. *RAM_o, IE, $SUPEX_o$, $SUBEX_o$ and DP1-DP4 collectively entail $WRAM^{\rightarrow}_{1}$*

Proof of Observation 6. We divide the proof into two parts: (I) a proof that if $A \in (K * A) * B$, then if $A > (B > C) \in K$, then $C \in (K * A) * B$, and (II) a proof that if $A \notin (K * A) * B$ and $B > C \in K * \neg A$, then if $A > (B > C) \in K$, then $C \in (K * A) * B$ (where $A, B, C \in L_o$).

Regarding (I): Assume $A \in (K * A) * B$ and $A > (B > C) \in K$ for conditional proof. From the latter, by the left-to-right half of IE, we have $A \wedge B > C \in K$. By RAM_o, this yields $C \in K * A \wedge B$. By DP1, we then recover $C \in (K * A) * A \wedge B$ and hence, by $SUPEX_o$, $C \in Cn((K * A) * B \cup \{A\})$. But by closure of belief sets and the fact that $A \in (K * A) * B$, $Cn((K * A) * B \cup \{A\})$ is itself equal to $(K * A) * B$. Hence $C \in (K * A) * B$, as required.

Regarding (II): We first establish two useful lemmas. The first lemma is:

Lemma 1: If $C \in (K * A) * B$, then either $C \in K * B$ or $C \in K * A \wedge B$

To see why this holds, assume $C \in (K * A) * B$, and hence (1) $C \in Cn((K * A) * B \cup \{\neg A\})$ and (2) $C \in Cn((K * A) * B \cup \{A\})$. We now consider two cases: $\neg A \in (K * A) * B$ and $\neg A \notin (K * A) * B$.

Assume $\neg A \in (K * A) * B$. It follows by consistency and closure of belief sets that $\neg\neg A \notin (K * A) * B$. From this, in conjunction with (1) and $SUBEX_o$, we recover $C \in (K * A) * \neg A \wedge B$. By DP2, since $\neg A \in Cn(\neg A \wedge B)$, it then follows that $C \in K * \neg A \wedge B$, and hence, by $SUPEX_o$, that (3) $C \in Cn(K * B \cup \{\neg A\})$. Furthermore, from the fact that $\neg A \in (K * A) * B$, we also have, by DP4, $\neg A \in K * B$, and hence $Cn(K * B \cup \{\neg A\}) = K * B$. The fact that $C \in K * B$ then immediately follows from (3) by the equality just established.

Assume that $\neg A \notin (K * A) * B$. It then follows from (2) and $SUBEX_o$ that $C \in (K * A) * A \wedge B$. By DP1, since $A \in Cn(A \wedge B)$, we then have $C \in K * A \wedge B$. This completes the proof of Lemma 1.

The second lemma is:

Lemma 2: If both $C \in K * B$ and $C \in K * A \wedge B$, then $C \in (K * A) * B$

 © Chandler 2016

To see why this holds, assume that both $C \in K * B$ and $C \in K * A \wedge B$. We now consider two cases: $A \in K * B$ and $A \notin K * B$.

Assume $A \in K * B$. It follows by DP3 that $A \in (K * A) * B$, and hence that $\mathrm{Cn}((K * A) * B \cup \{A\}) = (K * A) * B$. Since we have assumed that $C \in K * A \wedge B$, it therefore follows by SUPEX_o that $C \in (K * A) * B$, as required.

Assume $A \notin K * B$, and hence, by closure of belief sets, that $\neg\neg A \notin K * B$. Since $C \in K * B$, and hence $C \in \mathrm{Cn}(K * B \cup \{\neg A\})$, it then follows from SUBEX_o that $C \in K * \neg A \wedge B$. But we have also assumed $C \in K * A \wedge B$. Furthermore, $A \in \mathrm{Cn}(A \wedge B)$ and $\neg A \in \mathrm{Cn}(\neg A \wedge B)$. It therefore follows, by DP1 and DP2, that $C \in (K * A) * \neg A \wedge B \cap (K * A) * A \wedge B$. From SUPEX_o, we then recover the desired result that $C \in (K * A) * B$. This completes the proof of Lemma 2.

With all this in hand, assume $A > (B > C) \in K$ for conditional proof. The conjunction of the left-to-right half of IE with RAM_o then yields $C \in K * A \wedge B$. Now assume $B > C \in K * \neg A$, again for conditional proof. By the left-to-right direction of RAM_o, we have $C \in (K * \neg A) * B$. By Lemma 1 we therefore have either $C \in K * B$ or $C \in K * \neg A \wedge B$. Either way, we have $C \in K * B$, since the latter also follows from $C \in K * \neg A \wedge B$ by SUPEX_o and the fact that $C \in K * A \wedge B$. So both $C \in K * A \wedge B$ and $C \in K * B$. It therefore follows, by Lemma 2, that $C \in (K * A) * B$, as desired. □

Observation 7. *Given* INC_o, *SUC, WDP2 and ID, if, for pairwise inconsistent* $A, B, C \in L_o$, *we have* $\neg A, \neg B, \neg C \notin K$ *then (i) neither* $A \vee B \in ((K * A) * A \vee B) * \neg A$ *nor* $A \vee C \in ((K * A) * A \vee C) * \neg A$. *Furthermore, (ii) neither* $\neg A > B \in K * \neg(A \vee B)$ *nor* $\neg A > C \in K * \neg(A \vee C)$

Proof of Observation 7. Assume for conditional proof that, for pairwise inconsistent $A, B, C \in L_o$, we have $\neg A, \neg B, \neg C \notin K$.

Regarding (i): We simply prove that $A \vee B \notin ((K * A) * A \vee B) * \neg A$ (the proof that $A \vee C \in ((K * A) * A \vee C) * \neg A$ is analogous). Since both A and B are inconsistent with C, and $\neg C \notin K$, it follows by closure of belief sets that $A \vee B \notin K$.

Now assume for *reductio* that $A \vee B \in K * \neg A$. By INC_o, we have $A \vee B \in \mathrm{Cn}(K \cup \{\neg A\})$, and hence, by the deduction theorem, $\neg A \supset A \vee B \in K$. But, by closure of belief sets, this would mean that $A \vee B \in K$. Contradiction. Hence $A \vee B \notin K * \neg A$.

 © Chandler 2016

Assume for *reductio* again that $A \vee B \in ((K * A) * A \vee B) * \neg A$. Since $A \vee B \in \mathrm{Cn}(A)$, it follows, by SUC and closure of belief sets, that $A \vee B \in K * A$. By ID, we then recover $(K * A) * A \vee B = K * A$, and hence $A \vee B \in (K * A) * \neg A$. We then have, by WDP2, $A \vee B \in K * \neg A$. Contradiction. Hence $A \vee B \notin ((K * A) * A \vee B) * \neg A$.

Regarding (ii): We simply prove that $\neg A > B \notin K * \neg(A \vee B)$ (the proof that $\neg A > C \in K * \neg(A \vee C)$ is analogous). Assume $\neg A > B \in K * \neg(A \vee B)$ for *reductio*. By RAM_o, we have $B \in (K * \neg(A \vee B)) * \neg A$. Since $\neg(A \vee B)$ is consistent, given our initial assumption, we have, by SUC, $\neg(A \vee B) \in K * \neg(A \vee B)$, and hence, by closure of belief sets, $\neg A \in K * \neg(A \vee B)$. By ID, it then follows that $(K * \neg(A \vee B)) * \neg A = K * \neg(A \vee B)$. So, since $B \in (K * \neg(A \vee B)) * \neg A$, we have $B \in K * \neg(A \vee B)$, and hence, by closure and consistency of belief sets, $\neg(A \vee B) \notin K * \neg(A \vee B)$. Contradiction. □

Observation 8. *$SUPEX_o$, $SUBEX_o$, DP1, DP2, DP3 and RAM_o collectively entail that if $A \notin (K * A) * B$ and $B > C \notin K * \neg A$, then $B > C \notin K$*

Proof of Observation 8. We prove the logically equivalent proposition that if $A \notin (K * A) * B$ and $B > C \in K$, then $B > C \in K * \neg A$. Assume that $A \notin (K * A) * B$. From this fact, it follows by DP3 that $A \notin K * B$, and hence, by closure of belief sets, that $\neg\neg A \notin K * B$. Assume that $B > C \in K$. It follows by RAM_o that $C \in K * B$, and hence that $C \in \mathrm{Cn}(K * B \cup \{\neg A\})$. By SUBEX_o, we then have $C \in K * \neg A \wedge B$. From this, in conjunction with the fact that $C \in K * B$, it follows by Lemma 2 of Observation 6 that $C \in (K * \neg A) * B$, and hence, by RAM_o, that $B > C \in K * \neg A$, as required. □

Observation 9. *Given INC_o, if (i) $A \triangleright\!\mid B \in K$ iff $B \notin K * A$, then (ii) if $A \in K$ and $B \notin K$, then $A \triangleright\!\mid B \in K$*

Proof of Observation 9. Assume that $A \in K$. It follows by closure of belief sets that $K = \mathrm{Cn}(K \cup \{A\})$. Now further assume that $B \notin K$. It follows that $B \notin \mathrm{Cn}(K \cup \{A\})$. By the contrapositive of INC_o, we then have $B \notin K * A$, and hence, by the right-to-left half of (i), $A \triangleright\!\mid B \in K$ □

Observation 10. *SUC and ID jointly entail that, for non-contradictory $A, B \in L_o$, if (i) $A \triangleright\!\mid B \in K$ iff $B \notin (K * B) * A$ and (ii) if $A \in K$ and $B \notin K$, then $A \triangleright\!\mid B \in K$, then (iii) if $A \in K$ and $B \notin K$, then $A \notin K * B$.*

 © Chandler 2016

Proof of Observation 10. Assume (ii). On the assumption that (i), this is equivalent to the claim that if $A \in K$ and $B \notin K$, then $B \notin (K * B) * A$. Now assume that for some non-contradictory $A, B \in L_0$, we have $A \in K$ and $B \notin K$, and hence $B \notin (K * B) * A$. Now assume for *reductio* that $A \in K * B$. By ID, we have $(K * B) * A = K * B$. It follows from the fact that B is non-contradictory, in conjunction with SUC, that $B \in K * B$, and hence, by the previous equality, that $B \in (K * B) * A$. Contradiction. □

References

Adams, Ernest W. 1975: *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: Reidel.

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson 1985: 'On the Logic of Theory Change: Partial Meet Contraction and Revision Functions'. *Journal of Symbolic Logic*, 50(2), pp. 510–24.

Booth, Richard, and Thomas Meyer 2006: 'Admissible and Restrained Revision'. *Journal of Artificial Intelligence Research*, 26(1), pp. 127–51.

Boutilier, Craig 1996: 'Iterated Revision and Minimal Change of Conditional Beliefs'. *Journal of Philosophical Logic*, 25(3), pp. 263–305.

Bradley, Richard 2007: 'A Defence of the Ramsey Test'. *Mind*, 116, pp. 1–21.

—— 2012: 'Restricting Preservation: A Response to Hill'. *Mind*, 121, pp. 147–59.

Cantwell, John 1999: 'Some Logics of Iterated Belief Change'. *Studia Logica: An International Journal for Symbolic Logic*, 63(1), pp. 49–84.

Chalmers, David, and Alan Hájek 2007: 'Ramsey + Moore = God'. *Analysis*, 67(2), pp. 170–2.

Chandler, Jake 2013: 'Transmission Failure, AGM style'. *Erkenntnis*, 78(2), pp. 383–98.

Crocco, Gabriella, Luis Fariñas del Cerro, and Andreas Herzig (eds.) 1995: *Conditionals: From Philosophy to Computer Science*. Oxford: Clarendon Press.

Eells, Ellery, and Brian Skyrms (eds.) 1994: *Probability and Conditionals: Belief Revision and Rational Decision*. Cambridge: Cambridge University Press.

Etlin, David 2009: 'The Problem of Noncounterfactual Conditionals'. *Philosophy of Science*, 76(5), pp. 676–88.

 © Chandler 2016

Fuhrmann, André, and Hans Rott (eds.) 1996: *Logic, Action, and Information: Essays on Logic in Philosophy and Artificial Intelligence*. Berlin and New York: de Gruyter.

Gärdenfors, Peter 1979: 'Conditionals and Changes of Belief'. In Niiniluoto and Tuomela 1979, pp. 381–404.

——1986: 'Belief Revisions and the Ramsey Test for Conditionals'. *Philosophical Review*, 95(1), pp. 81–93.

——1987: 'Variations on the Ramsey Test: More Triviality Results'. *Studia Logica*, 46(4), pp. 321–7.

Hájek, Alan, and Ned Hall 1994: 'The Hypothesis of the Conditional Construal of Conditional Probability'. In Eells and Skyrms 1994, pp. 75–113.

Hansson, Sven Ove 1992: 'In Defense of the Ramsey Test'. *Journal of Philosophy*, 89(10), pp. 499–521.

Jin, Yi, and Michael Thielscher 2007: 'Iterated Belief Revision, Revised'. *Artificial Intelligence*, 171(1), pp. 1–18.

Leitgeb, Hannes 2011: 'God – Moore = Ramsey (A Reply to Chalmers and Hájek)'. *Topoi*, 30(1), pp. 47–51.

Lewis, David 1976: 'Probabilities of Conditionals and Conditional Probabilities'. *Philosophical Review*, 85(3), pp. 297–315.

Lindström, Sten, and Włodzimierz Rabinowicz 1992: 'Belief Revision, Epistemic Conditionals and the Ramsey Test'. *Synthèse*, 91(3), pp. 195–237.

——1995: 'The Ramsey Test Revisited'. In Crocco et al. 1995, pp. 131–82.

McGee, Vann 1985: 'A Counterexample to Modus Ponens'. *Journal of Philosophy*, 82(9), pp. 462–71.

Nayak, Abhaya C. 1994: 'Iterated Belief Change Based on Epistemic Entrenchment'. *Erkenntnis*, 41(3), pp. 353–90.

Niiniluoto, Ikka, and Raimo Tuomela (eds.) 1979: *The Logic and Epistemology of Scientific Change. Acta Philosophica Fennica*, 30(2–4). Amsterdam: North-Holland.

Rabinowicz, Włodzimierz 1996: 'Stable Revision, or Is Preservation Worth Preserving?' In Fuhrmann and Rott 1996, pp. 101–28.

Rott, Hans 1986: 'Ifs, Though, and Because'. *Erkenntnis*, 25(3), pp. 345–70.

——1989: 'Conditionals and Theory Change: Revisions, Expansions, and Additions'. *Synthèse*, 81(1), pp. 91–113.

——2011: 'Reapproaching Ramsey: Conditionals and Iterated Belief Change in the Spirit of AGM'. *Journal of Philosophical Logic*, 40(2), pp. 155–91.

 © Chandler 2016

Supporters and Underminers: Reply to Chandler

RICHARD BRADLEY
London School of Economics and Political Science
r.bradley@lse.ac.uk

I argue that two of the standard axioms of the AGM theory of belief revision stand in the way of it serving as the basis for an adequate account of defeasible reasoning, because they respectively disallow the adoption of beliefs not logically entailed by those previously learned and the abandonment of those not contradicted by them.

1. Introduction

In 'A Defence of the Ramsey Test', I argued that the triviality results for the Ramsey Test hypothesis were misdirected and that we should instead reject one or more of the premises of these results. The Preservation condition, in particular, seemed doubtful to me and I offered two arguments for rejecting it; one which relied on a commutativity condition for belief revision and the other on a general version of Modus Ponens. In his discussion paper Chandler contests the significance of these results, though not the broader argument in favour of the Ramsey Test hypothesis. Since I concur that neither the commutativity condition nor Modus Ponens is incontestable and Chandler does not, I think, deny the need to restrict the Preservation condition, I won't have much to say about this part of his paper. Instead I will focus my remarks on the more positive parts of his paper and on the general issue of the suitability of the AGM framework for modelling defeasible reasoning.

2. Belief preservation

In my reply to Brian Hill's (2011) critical notice, I argued that we do not want a version of the Preservation condition that forces us to

Mind, Vol. 126 . 502 . April 2017
© Bradley 2016
doi:10.1093/mind/fzw003 Advance Access publication 7 December 2016

retain beliefs that have been undermined by what we have learnt. If I initially believe that *B* but my reasons for doing so are undermined by learning that *A*, I should be free to give up my beliefs in *B* even if my beliefs, together with *A*, do not logically entail the falsity of *B*. Suppose, for instance, that I measure the temperature of a glass of water using a thermometer and, as a result, come to believe that it is 15 degrees Celsius, but that I am subsequently told that the thermometer has not been functioning correctly for some time. I should be free to abandon my belief about the temperature of the water even if I did not explicitly hold the prior belief that the thermometer was functioning correctly. For I cannot reasonably be expected to anticipate all the things that I might learn that would undermine my current belief were I to learn that they were true. But this is precisely what the unrestricted Preservation condition requires.

The phenomenon of belief undermining speaks in favour of the Acceptance interpretation of the AGM model and against the Informational interpretation. On the Acceptance interpretation, an epistemic state is just the set of propositions that the agent accepts, perhaps because they are sufficiently credible or because they constitute working hypotheses or because they are central to her plans. On the Informational interpretation, on the other hand, the agent's epistemic state is the set of propositions or sentences that the agent knows to be true or is certain about. This latter interpretation leaves little room for the possibility of learning something that undermines one's beliefs, restricting it to cases in which one learns something that contradicts what one believes (which should be a rare event if one's beliefs are certainties).

On the Informational interpretation, beliefs should be preserved unless contradicted, precisely because they are certainties. On the Acceptance interpretation, on the other hand, the motivation for preserving beliefs is an economic one. The thought is that it is costly to acquire beliefs and that consequently we should not give them up without good reason. Learning something inconsistent with what one believes is clearly such a reason, but so too is learning something that undermines what one currently believes, for instance by casting doubt on evidence which supports a particular belief. So the Acceptance interpretation would mandate weakening the Preservation condition to require that upon learning something I should retain any current beliefs not undermined by what I have learnt. In this spirit, let us say that a belief *B* is stable under revision by some proposition *A* of the belief set in which it is contained, just in

 © Bradley 2016

case A does not undermine B. Then the principle I proposed can be formulated as:

> (Stable Belief Preservation) Suppose that $B \in K$, $\neg A \notin K$ and A does not undermine B in K. Then $B \in K * A$.

This formulation was vague on the question of what it takes for one belief to undermine another. Intentionally so, because although a belief being inconsistent with another is surely sufficient for it to undermine it, there seems to be no consensus amongst epistemologists as to how to characterise this undermining relation more precisely. I did however argue that a conditional could be undermined by acceptance of a proposition that was inconsistent with its antecedent and, on that basis, tentatively suggested something close to what Chandler labels PRES_2^*.

Chandler argues that PRES_2^* is too strong, offering both a counterexample and a technical result to support this claim (Observation 6). Chandler's counterexample is intuitively compelling but assumes a concept of defeasibility that I find doubtful.[1] (More on this later.) Nonetheless I think Chandler is right to say that PRES_2^* is too strong. Here is an alternative counterexample to it that also serves to illustrate his technical result. Suppose that of the horses running this afternoon, I consider only Speedy, Steady and Plodder to be potential winners. So I believe that if Steady is not going to win, then Plodder will win if Speedy (for some reason) does not. I then learn that Plodder has been disqualified. Should I continue to believe that if Steady doesn't win, then Plodder will if Speedy doesn't? Of course not, for I now know that Plodder can't win under any circumstances. But PRES_2^* licenses preserving this belief since Plodder's disqualification is logically consistent with the supposition that Speedy won't win. So what PRES_2^* fails to pick up is that the disqualification undermines my belief that Plodder will win if Speedy does not. The upshot is that PRES_2^* is not the right way to make (Stable Belief Preservation) more precise.

[1] Chandler supposes (i) agnosticism about the truth of factual propositions A, B, C and D, (ii) that A is a 'sufficient but defeasible reason' to believe B and (iii) that $B \in K * A$. But from (i) and (iii) it follows by VAC_o and INC_o that $B \in Cn(K \cup \{A\})$. And this, in my view, contradicts (ii) since A cannot be a defeasible reason to believe B given K, when they jointly imply it. But B is a non-defeasible reason to believe D, so C cannot undercut D as his example presumes.

 © Bradley 2016

3. Defeasible reasoning

One belief undermines another, Chandler suggests, if, having committed to the truth of the latter, the supposition that the former is true leads to this commitment disappearing. More formally, he postulates that $A \not\triangleright B \in K \Leftrightarrow B \notin (K * B) * A$, where $A \not\triangleright B$ reads as 'A is an underminer for B'. For our purposes however it is more natural to treat undermining as a relation between beliefs within a particular belief set (thus as a ternary relation on belief sets and pairs of sentences contained in them). So, in the spirit of Chandler's proposal, I suggest the following characterisation of belief undermining:

> (Undermining) If $\neg A \notin K$, then A undermines B in K iff $B \notin (K * B) * A$

Note the restriction to cases in which $\neg A$ does not belong to K. This is required because when $\neg A$ does belong to K, revision by A undermines K as a whole rather than B alone.

Undermining, so characterised, may reasonably be regarded as the opposite side of the same coin to (Stable Belief Preservation). While the latter attempts to constrain belief preservation by drawing on an unspecified notion of undermining, (Undermining) draws on an unspecified belief revision operation to characterise the notion of belief undermining. Jointly they give us what I think we should regard as a basic principle (even platitude) relating conservative belief revision to defeasible reasoning:

> (DEF) If $\neg A \notin K$ and $B \in K$, then $B \in K * A$ iff A does not undermine B in K

How far does this take us? One might hope that, together with the AGM postulates, it will give us the precise characterisation of undermining required, and hence of the restriction that should be imposed on Preservation. But alas not, for VAC_o robs (Undermining) of its interest. When A is consistent with K and B is in K, VAC_o implies that $B \in K * A$. So A can undermine beliefs in K only when A is inconsistent with K. It follows that if (DEF) is to serve as a regulative principle of defeasible reasoning then VAC must be given up, even for factual sentences. And without VAC the vagueness in (Stable Belief Preservation) remains.

Defeasible reasoning is ampliative: it produces conclusions that are not deductive consequences of the propositions we accept. For instance, when we infer properties of a population from that of a sample, or infer a universal generalisation on the basis of a finite

 © Bradley 2016

number of observed instances of it, we add content to our beliefs beyond that contained in what we have observed. Amplification is just the flip-side of undermining. In cases of the former we take on beliefs not logically entailed by those already accepted; in cases of the latter we give up beliefs that are logically consistent with those we accept. A theory of defeasible reasoning should allow for both.

In the final section of his critical notice, Chandler makes a suggestion in this regard that I find helpful. Amplification has its source in the possibility of a belief being supported by one or more other beliefs, even when the latter don't entail the former, i.e. in the possibility of one or more beliefs being a non-deductive reason to accept another. Chandler proposes that the belief that A is a reason to believe that B just in case, having withdrawn commitment to B, the supposition that A leads to the adoption of the belief in B. In this spirit, but again treating it as a relation between beliefs and a belief set, we can say:

(Supporting) A supports B in K iff $B \in (K \dot{-} B)*A$

The relation of supporting, so defined, is in tension with INC. For INC says that if you accept B upon revising your beliefs by A, then B must be a logical consequence of A together with your initial set of beliefs. So if your newly accepted belief B is not a deductive consequence of your other beliefs, it must have been obtained by reasoning that violates INC. So it seems that *just as the mark of undermining is the failure of VAC, the mark of amplification is the failure of INC.*[2]

Let's take stock. I think we should take both (Undermining) and (Supporting) as regulative of a theory that links defeasible reasoning and belief revision. Different candidates for such a theory will be more or less conservative in the degree to which they allow for beliefs to be adopted and rejected. On an Anglo-Saxon 'hire and fire' model of belief revision, beliefs will be readily accepted, and equally readily dispensed with, as experience or deliberation produces reasons for and against them. On a more Continental model, the barriers to entry into a belief set will be higher and those that are already in it correspondingly more entrenched. These differences will be reflected in the balance that the models strike between the need to accept truths and to reject falsehoods, a balance achieved by calibrating the conditions on belief revision, and in particular those for belief inclusion and exclusion, to the adopted consequence relation.

[2] Levi (1996) makes a similar point and backs it up with a good deal more detailed analysis than I can give here.

 © Bradley 2016

The combination of a supra-classical consequence relation together with INC and VAC presented in Gärdenfors (1988) gives a system of reasoning that is more of the Continental than Anglo-Saxon variety. What we have learnt from the debate following my 2007 paper, and in particular from Chandler's contribution, is that this system does not leave enough space for defeasible reasoning. Not just because it is inconsistent with the Ramsey Test hypothesis but because it reduces the relations of undermining and support between beliefs to cases of logical inconsistency and logical entailment. The cost of such conservatism is two-fold: restricted opportunities to adopt useful beliefs and the burden of poor, but entrenched, beliefs still awaiting decisive refutation.[3]

References

Bradley, Richard 2007: 'A Defence of the Ramsey Test'. *Mind*, 116, pp. 1–21.

——2012: 'Restricting Preservation: A Response to Hill'. *Mind*, 121, pp. 147–59.

Gärdenfors, Peter 1988: *Knowledge in Flux*. Cambridge, MA: MIT Press.

Hill, Brian 2011: 'Defending the Ramsey Test: What is wrong with Preservation?' *Mind*, 121, pp. 131–46.

Levi, Isaac 1996: *For the Sake of Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning*. Cambridge: Cambridge University Press.

[3] I am grateful to Jake Chandler for his useful comments on an earlier draft.

 © Bradley 2016

Preservation and Postulation: Lessons from the New Debate on the Ramsey Test

Hans Rott
University of Regensburg
hans.rott@ur.de

Richard Bradley has initiated a new debate, with Brian Hill and Jake Chandler as further participants, about the implications of a number of so-called triviality results surrounding the Ramsey test for conditionals. I comment on this debate and argue that 'Inclusion' and 'Preservation', which were originally introduced as postulates for the rational revision of factual beliefs, have little to recommend them in the first place when extended to languages containing conditionals. I question the philosophical method of postulation that was applied in the new debate, and instead base my arguments on explicit representations of belief states and canonical constructions of belief state revisions.

1. Introduction

Peter Gärdenfors (1978) used a novel theory of belief revision in order to come up with a logic for conditionals with the help of the so-called Ramsey test for conditionals. This theory of belief revision was developed in full detail in a celebrated paper by Alchourrón, Gärdenfors and Makinson (1985). In what may be regarded as an ironic twist in the research history, it was Gärdenfors himself who showed, only one year later, that the perfected 'AGM theory' of belief revision was not suitable for its original purpose. This is the upshot of Gärdenfors's (1986) famous 'triviality theorem'.

A new debate has recently arisen in this journal concerning the lessons that are to be drawn from the 'triviality results'. The new debate was initiated by Richard Bradley (2007), who added two more triviality results. Bradley's paper provoked responses by Brian Hill (2012) and Jake Chandler (2017), to which Bradley (2012; 2017) in turn replied. In this discussion note, I concur with the participants of the new debate that the Ramsey test is worth defending and that 'Inclusion' and 'Preservation' should be questioned. But I will argue that Inclusion and Preservation, which were originally introduced as postulates for the rational revision of factual beliefs, have little to

Mind, Vol. 126 . 502 . April 2017
doi:10.1093/mind/fzw028 Advance Access publication 1 March 2017
© Rott 2017

recommend them in the first place when extended to languages containing conditionals. I will question the philosophical method of postulation that was applied in the new debate, and instead base my arguments on explicit representations of belief states and canonical constructions of belief state revisions.

2. Belief sets, belief states, and belief change models

Consider a propositional language with the usual truth-functional operators $\neg$, $\wedge$ and $\vee$, as well as a binary conditional operator $>$.[1] The conditional-free sentences of the language constitute its *factual* fragment. I assume that the language contains right-nested conditionals of the form $A > B$, $A > (B > C)$, $A > (B > (C > D))$, and so on, where the constituent sentences, A, B, C, D, ..., are all factual. The language is not supposed to contain more problematic constructions like negations or other truth-functional compounds of conditionals which many have taken to presuppose that conditionals express propositions. Nor will I refer to left-nested conditionals.

As we shall see, right-nested conditionals can easily be interpreted by the Ramsey test based on models for iterated belief revision. I shall not, however, presume that the Ramsey test furnishes an analysis of natural language conditionals that invariably come with tense and mood. Since Adams (1970), indicative and subjunctive (open and counterfactual) conditionals have generally been thought to have different meanings. Advocates of the Ramsey test may hope that this difference can be captured by different ways of revising one's beliefs. But neither Bradley, Hill nor Chandler takes up such issues, and I shall not undertake to address this matter in my discussion. Let us take it that '$>$' expresses *Ramsey test conditionals*. So by definition, the Ramsey test will not be called into question in this paper. I leave it open here how close in meaning Ramsey test conditionals are to ordinary 'if ... then' constructions.

A regimented language is supposed to be governed by a logic reflecting what follows from what. One thing that has been overlooked in the new debate, I think, is that the interplay between logic and belief revision gives rise to a chicken-and-egg problem. On the one hand, seven out of eight AGM postulates for belief revision refer to a background logic, Cn, and in this sense logic seems to be prior to belief

[1] Bradley and Hill use the symbol '$\rightarrow$'. I adopt Chandler's notation here in order to avoid any danger that the conditionals discussed here are confused with material conditionals.

 © Rott 2017

revision. On the other hand, Gärdenfors used the Ramsey test for deriving a logic of conditionals from the properties of belief revision models. In this sense, belief revision appears to be prior to logic. When one requires that belief sets be 'consistent' and 'deductively closed', one should be maximally clear what these terms are supposed to mean. To what logic do they refer? Since our language contains conditionals, it is natural to assume that its logic should include some principles involving conditionals. In the new debate on the Ramsey test, the logical principles of Conditional Contradiction[2] and Modus Ponens[3] for $>$ are discussed. It is evident that this cannot amount to a full characterization of the conditional logic in question, but I will argue that it may be wiser not to start by using a conditional logic for the characterization of belief sets.

In AGM's original theory, doxastic states were identified with so-called *belief sets*, that is, sets of factual sentences closed under *Cn*.[4] But it was soon noticed that for iterated belief changes one needs changes of more complex *belief states* (also known as *doxastic states*). I shall adopt a qualitative approach and represent a belief state by a total preorder of possible worlds, where a *possible world* is simply identified here with an interpretation of the factual language.[5] Let W denote the set of all interpretations of the factual language, and $[A]$ denote the set of worlds at which the factual sentence A is true. To keep things mathematically simple, I shall assume that all such sets are finite. I use a non-strict relation $\preceq$ between possible worlds, and require, for

[2] Somewhat confusingly, each paper refers to a different version of Conditional Contradiction (CC): see Bradley (2007, p. 5), Hill (2012, p. 135), Bradley (2012, p. 148) and Chandler (2017, p. 582). That Bradley's first version cannot be right was pointed out by Hill and by Rott (2011, p. 161). Bradley's second version is still incorrect, probably due to a distorting typo. Chandler's version is not intended to be a logical, but a 'doxastic', principle.

[3] Bradley (2007, p. 8) introduces unrestricted Modus Ponens for $>$, Hill (2012, p. 136) brings into play another unrestricted version, Bradley (2012, p. 152) rejects both versions, and Chandler (2017, pp. 579, 591–595) discusses a restricted 'doxastic analogue' of Modus Ponens.

[4] *Pace* Bradley (2012, p. 154; 2017, p. 604), I think that the 'acceptance interpretation' of the belief change theories in the tradition of AGM is the only correct interpretation, and that the 'informational interpretation' is wrong, because these theories never talked about knowledge or certainty. The title of Gärdenfors (1988) is just a misnomer.

[5] Chandler (2017, pp. 579–580) identifies belief states with belief sets in languages containing conditionals. Given the Ramsey test, this is equivalent to the models used below. A principle that Chandler repeatedly uses in his proofs, namely,

(†) If $K = K'$, then $K*A = K'*A$

for the full language indeed depends for its motivation on the Ramsey test. However, Chandler (2017, pp. 591–592) seems to favour a solution that restricts the Ramsey test.

 © Rott 2017

the purposes of this paper, that it be a total preorder, that is, transitive and connected, over its field $F_{\preceq} = \{w \in W$: $v \preceq w$ or $w \preceq v$ for some $v \in W\}$. Intuitively, the relation $\preceq$ represents the *comparative plausibility* of possible worlds as seen from a person's doxastic point of view. The sentence '$v \preceq w$' expresses the fact that w is at least as plausible (or regarded as at least as likely to be the actual world) as v. The field $F_{\preceq}$, or briefly F when there is no danger of confusion, specifies the set of worlds that are *doxastically possible* for a person in belief state $\mathcal{B}$ represented by $\preceq$, in other words, the set of worlds that could possibly be the actual world, as far as the person's belief state is concerned. F is a subset of W; it need not be identical with W.

Two designated belief states are the 'blank slate' $\mathcal{O} = W \times W$ that represents complete doxastic emptiness (and accepts only logical truths) and the 'absurd belief state' $\mathcal{A} = \varnothing$ that represents complete doxastic confusion (and accepts everything).

A world w is *doxastically preferred* in the state $\mathcal{B}$ represented by $\preceq$ if w is maximal in F under $\preceq$, that is, if w is in F and there is no w' in F such that $w \prec w'$. (We write '$v \prec w$' for '$v \preceq w$ and not $w \preceq v$'.) The set $s^f(\mathcal{B})$ of factual sentences supported by the belief state $\mathcal{B}$, or the *factual belief set associated with* $\mathcal{B}$, is the set of sentences that are true at all doxastically preferred worlds. Call a belief state $\mathcal{B}$ *dogmatic* if it has only one doxastically preferred world and *undogmatic* otherwise. Dogmatic belief states support either A or $\neg A$, for every factual sentence A. Only notorious know-it-alls have an opinion about every single factual proposition. Reasonable persons suspend judgement about some propositions.

We now turn to the revision $\mathcal{B} * A = \preceq^*_A$ of a belief state $\mathcal{B} = \preceq$ by a factual sentence A. An important semantic idea is what may be called the *Maximal Rule*:

(MR) $\quad \max_{\preceq^*_A}(F^*_A) = \max_{\preceq}(F \cap [A])$[6]

The Maximal Rule says that the maximally plausible worlds in the revised belief state are the maximally plausible A-worlds in the prior state. Now we are prepared to formulate the *Ramsey test* for conditionals:

(RT) $\quad \mathcal{B}$ supports $A > B$ iff $\mathcal{B} * A$ supports B

[6] Notice that 'maximal' worlds with respect to $\preceq$ are always meant to be worlds from $F_{\preceq}$ and that $\max_{\preceq}(\varnothing) = \varnothing$. For more on the Maximal Rule, see Rott (2011, p. 169), where it is called 'the (AGM) constraint'.

 © Rott 2017

Let A, B and C be factual sentences. Then the set $s^c(\mathcal{B})$ of conditionals supported by the belief state $\mathcal{B}$ is determined as follows: $\mathcal{B}$ supports $A > B$ iff (by RT) $\mathcal{B} * A$ supports B iff (by the rule for factual belief) $\max_{\preceq_A^*}(F_A^*) \subseteq [B]$ iff (by MR) $\max_{\preceq}(F \cap [A]) \subseteq [B]$. Moreover, we find that $\mathcal{B}$ supports $A > (B > C)$ iff (by RT) $\mathcal{B} * A$ supports $B > C$ iff (by the result just obtained) $\max_{\preceq_A^*}(F_A^* \cap [B]) \subseteq [C]$. The set of sentences supported by the belief state $\mathcal{B}$, or the *belief set associated with* $\mathcal{B}$, is $s(\mathcal{B}) = s^f(\mathcal{B}) \cup s^c(\mathcal{B})$.[7] To give an example, suppose that Alice's belief state is chacterized by a plausibility preorder $\preceq$ such that in all doxastically preferred worlds, 'Douala is the capital of Cameroon' is true, and that in all maximal worlds in which this sentence is false, 'Yaoundé is the capital of Cameroon' is instead true. Then Alice's belief set includes the sentences 'Douala is the capital of Cameroon' and 'If Douala is (or were) not the capital of Cameroon, then it is (or would be) Yaoundé'.

The following concept will be needed later. A belief state refines another belief state if it keeps all the plausibility distinctions recognized by the latter. Refinements can elegantly be defined in terms of total preorders. The state $\mathcal{B}' = \preceq'$ is a *refinement* of the state $\mathcal{B} = \preceq$ (within a common subfield $V \subseteq F \cap F'$) iff for all worlds v and w (in V), if $v \preceq' w$ then $v \preceq w$.

Belief sets clearly do not capture everything that is relevant in the belief states of human agents. Even a plausibility preorder $\preceq$, a more complex non-propositional structure, is unlikely to represent all important features of a real belief state. But the idea here is that such a structure captures those aspects that are relevant for the evaluation of conditionals. Based on this idea, I will take the liberty of saying that a belief state is identical with (rather than just represented by) a plausibility preorder $\preceq$.

According to the Ramsey test analysis, a belief state $\mathcal{B}$ supports conditionals only relative to some revision function. To be more precise, let us introduce a *belief revision model* as a pair $\langle \mathbb{B}, * \rangle$, with $\mathbb{B}$ being a set of belief states and $*$ being a revision function that assigns to each pair consisting of a ('prior') belief state $\mathcal{B}$ and a factual sentence A another ('posterior') belief state $\mathcal{B} * A$. Now we say that a sentence A is a *consequence* of a set of sentences H—in symbols, $A \in Cn(H)$—if and only if for any belief state $\mathcal{B}$ in any belief revision model $\langle \mathbb{B}, * \rangle$, if $\mathcal{B}$ supports all elements of H, then it supports A too. This condition can be read in two directions. Given a certain logic Cn,

[7] This concept of support and the s-function are borrowed from Hansson (1992, p. 525).

 © Rott 2017

the left-to-right direction is just the first AGM postulate, called Closure. However, given the notion of a belief change model, with its associated notion of a belief state, the right-to-left direction may be used to establish a logic for conditionals.[8] A sentence A is consistent with a set of sentences H if and only if there is a belief state in some belief revision model $\langle \mathbb{B}, * \rangle$ that supports $H \cup \{A\}$.

So I suggest that the solution of the chicken-and-egg problem mentioned above proceeds in three steps. First, use a base logic Cn^f for the factual language to define possible worlds. Then include some non-propositional structure (plausibility preorders) in the possible-worlds representation of belief states and their dynamics. Finally, employ belief revision models to characterize the logic Cn of the full language with right-nested conditionals.

3. Revising by compatible information: Inclusion and Preservation

The third and the fourth AGM postulates relate the revision function to the set K of currently held factual beliefs, and express principles of *conservatism* or *minimal change* for the case in which the input is logically compatible with the belief set K.

(Inclusion) If $B \in K*A$, then $B \in Cn(K \cup \{A\})$

(Preservation) If $\neg A \notin K$ and $B \in K$, then $B \in K*A$

These postulates present substantial recommendations for how to perform revisions by input sentences A that are consistent with the prior beliefs in K. Inclusion states that the agent should not acquire more beliefs than are required by the first two AGM postulates, Closure and Success. Preservation tells her not to give up more beliefs than are required by the fifth AGM postulate, Consistency.[9] Inclusion and

[8] For more details on this proposal, see the 'belief state logic' Cn^- in Rott (2011, especially p. 165).

[9] For the list of the eight AGM postulates for belief revision in their canonical numbering, see Gärdenfors (1988, pp. 54–6) or Bradley (2007, p. 20). AGM actually have as the fourth postulate a restricted converse of Inclusion, namely,

(Vacuity) If $\neg A \notin K$, then $Cn(K \cup \{A\}) \subseteq K*A$

The additional strength of Vacuity over Preservation can be gained from Closure and Success. In order to avoid redundancies in the axiomatization, it is advisable to use the simpler and weaker postulate of Preservation. Note incidentally that 'Inclusion' is not a good name because many of the AGM postulates are about inclusion relations. 'Vacuity' for the fourth AGM

 © Rott 2017

Preservation can also be said to encode an idea of diachronic coherence requiring that the prior and the posterior belief states (or more generally, the members in a sequence of belief states) somehow 'hang together'. The idea is restricted here to factual inputs that are logically compatible (i.e. consistent) with the current belief set. In the context of the semantic modelling of belief states, K and $K*A$ stand for $s(\mathcal{B})$ and $s(\mathcal{B}*A)$, respectively. The question now is whether Inclusion and Preservation are still valid if the language contains Ramsey test conditionals.

4. Postulation: against Inclusion and Preservation of conditionals

The postulate of Preservation has long been one of the prime suspects for the triviality results associated with the Ramsey test (Rott 1989, Hansson 1992, Boutilier 1994). In particular, it has been argued that if the object language contains 'autoepistemic' operators or conditionals, it is a mistake to assume that the revision of a belief set K by (factual) information A consistent with K is identical with its 'expansion' $Cn(K \cup \{A\})$.[10]

Preservation has been the most salient object of criticism in the new debate, too. However, the result is not very satisfying. The central suggestions for how to weaken Preservation offered by Hill (2012, p. 140) and Bradley (2012, pp. 157–8) are very complex, and do not have much intuitive appeal.[11] Chandler's charitable rendering of Bradley's intention (Chandler 2017, p. 589, condition $PRES_2^*$) looks much more agreeable, but both Chandler and Bradley now think that it cannot be right.

Chandler (2017) lists, in display mode, around twenty relevant principles for belief change. In addition, there are quite a number of further principles discussed by Bradley and Hill. Many of the less simple

---

postulate is an equally infelicitous name: it is unclear what is supposed to be vacuous here. I use these labels for the sake of continuity.

[10] Detailed overviews on the state of the art as of around 2000 concerning the Ramsey test are given by Lindström and Rabinowicz (1998) and Nute and Cross (2001, pp. 46–85).

[11] Hill and Bradley make essential use of negated conditionals, the semantics of which are very controversial. This makes it hard, in my opinion, to accept their arguments. (The present paper stays away from the problems surrounding the 'Negative Ramsey Test'; cf. Bradley 2007, pp. 6–7; 2012, p. 157; and Hill 2012, pp. 143–4.) Hill draws extensively on other truth-functional compounds of conditionals as well; but it is unclear how to evaluate them using the Ramsey test.

 © Rott 2017

ones bear the marks of ad hocness. After so much finessing and varying of the main postulates, I find it hard to trust my own or anyone else's intuitions about them. It seems more enlightening to me to see which principles are satisfied by which of the most natural constructive models for belief revisions. Thus a general methodological point may constitute the most important lesson of the new debate. In an early and insightful paper, David Makinson (1985, p. 350) referred to Bertrand Russell's famous quip, 'The method of "postulating" what we want has many advantages; they are the same as the advantages of theft over honest toil' (Russell 1919, p. 71). In the case of belief change, the toil consists in the definition of constructions of belief change operations that may actually have the properties that we desire.

AGM provided three different constructions for their postulates, namely, partial meet revision, safe revision, and entrenchment-based revision. I will focus here on a possible-worlds semantics which is very closely related to partial meet revision.[12] While AGM themselves said very little about iterated changes of beliefs, other researchers went on to extend their account. Based on the representation of belief states by total preorders of possible worlds (see §2), four *canonical constructions* have been advocated since the 1990s: moderate revision, conservative revision, radical revision, and restrained revision.[13] These are the most straightforwardly reasonable ways to extend AGM belief revision to the iterated case. The constructions have a clear motivation, and they allow precise answers as to which variants of the principles discussed by Bradley, Hill and Chandler—Inclusion, Preservation, Commutativity, and doxastic Modus Ponens—are valid for which construction. They can be regarded as providing a semantics for the postulates of belief revision theory.

The canonical constructions form two clearly separable groups. Conservative and restrained revision both satisfy Flat Conditionals Preservation:

(FCP) Suppose that A, B, C are factual and $\neg A \notin K$; then

if $B > C \in K$, then $B > C \in K * A$

[12] This was made clear by Grove (1988) and Katsuno and Mendelzon (1991). Results concerning the constructions of the corresponding belief contraction functions are conveniently surveyed in Gärdenfors (1988, ch. 4) and Gärdenfors and Rott (1995, pp. 61–81).

[13] For the definitions, see the appendix to this paper.

 © Rott 2017

but they do not satisfy Consistent Import-Export:

(CIE) Suppose that A, B, C are factual and $\neg A \notin Cn(B)$; then

$C \in (K*A)*B$ if and only if $C \in K*(A \wedge B)$

For moderate and radical revision, exactly the opposite is true. This suggests that FCP and CIE are incompatible. It turns out that indeed they are, on pain of triviality, if we embrace the very basic AGM postulates of Success, Closure, Consistency, and Preservation of factual sentences. Given the Ramsey test, FCP and CIE jointly entail

(‡) Suppose that A, B, C are factual, $\neg A \notin K$, and $\neg A \notin Cn(B)$; then

if $C \in K*B$, then $C \in K*(A \wedge B)$

This is surely an unreasonable property. Consider two (factual) propositional atoms p and q, let W be the set of four interpretations over p and q, and define the belief state $\mathcal{B}_1 = \preceq_1$ by $v \preceq_1 w$ iff $v \in [p \wedge q]$ or $w \notin [p \wedge q]$. Note that $F_1 = W$, that $\preceq_1$ is transitive and connected, and that the set of most plausible worlds is $[\neg(p \wedge q)]$. Clearly, $\neg p$ is not in $K = s^f(\mathcal{B}_1)$ and not in $Cn(q)$. But by Success, Preservation of factual sentences, and Closure, $\neg p$ is in $K*q = s^f(\mathcal{B}_1*q)$, while by Success and Consistency, $\neg p$ is not in $K*(p \wedge q) = s^f(\mathcal{B}_1*(p \wedge q))$, contradicting (‡).

So FCP and CIE are incompatible, if the revision function conforms to the fairly innocuous AGM postulates mentioned and if $\mathcal{B}_1$ indeed represents a belief state. The latter condition is a semantic one, and its role is similar to that of the non-triviality assumptions that are used in the known so-called triviality results based on the method of postulation. Still, I want to urge that it would be wrong to call the observation we just made a 'triviality result'. There is nothing trivial about two desiderata on a wish list that are not simultaneously satisfiable.

Since the publication of Gärdenfors's triviality result, it has been known that *some* of the AGM postulates must be invalid if the language contains conditionals that go by the Ramsey test. But what is the situation concerning the particular postulates that are most interesting to us, Inclusion and Preservation, once we have adopted the constructive point of view?

Turning to Inclusion first, we need to consider the set of consequences of $s(\mathcal{B}) \cup \{A\}$. Inclusion holds for all canonical constructions if the language is purely factual. But for the full language, if a belief

 © Rott 2017

state $\mathcal{B}$ does not already support a factual sentence A, then A cannot naturally be added to the belief set associated with $\mathcal{B}$.[14] Only a dogmatic belief state $\mathcal{B}'$ can support the sentences supported by $\mathcal{B}$ as well as A, and revising into dogmatism is normally ruled out by Inclusion for factual sentences. Thus, in all but a number of very special cases, the set $s(\mathcal{B}) \cup \{A\}$ is the inconsistent set of all sentences. This makes Inclusion trivially true.[15] Interestingly, however, Inclusion (non-trivially) fails for all canonical constructions if our language only allows *flat* Ramsey test conditionals. To see this, consider the blank slate $\mathcal{O} = W \times W$. The revised belief state $\mathcal{O}*(\neg p \vee q)$ supports the conditional $p > q$, for all canonical revision functions $*$. The belief state $\mathcal{B}_2 = \preceq_2$ defined by $v \preceq_2 w$ iff $v \in [p]$ or $w \notin [p]$ supports $s(\mathcal{O})$ and $\neg p \vee q$.[16] But $\mathcal{B}_2$ fails to support the conditional $p > q$, for all canonical revision functions $*$. So there is a belief state supporting $s(\mathcal{O}) \cup \{\neg p \vee q\}$ but not $p > q$, hence the latter is not a consequence of the former. Together with the fact that $p > q$ is in $s(\mathcal{O}*(\neg p \vee q))$, this contradicts Inclusion. We conclude that if the language contains Ramsey test conditionals, then Inclusion is either false (in the restricted case without nestings) or trivially true (if right-nesting is allowed).

Moving on to Preservation, it is clear that all canonical constructions satisfy Preservation for a purely factual language. Concerning Preservation of flat conditionals, we first note that conservative and restrained revision, but not moderate revision, are *structure-preserving* in the following sense: if a factual sentence A is consistent with $\mathcal{B}$, then $\mathcal{B}*A$ refines $\mathcal{B}$. In radical revision (which operates only on the first component of $\mathcal{B}$), $\mathcal{B}*A$ refines $\mathcal{B}$ within the restricted range $W \cap [A]$, but cutting off the worlds outside $W \cap [A]$ has an effect similar to the effect of downgrading them with respect to their plausibility. From this it follows that conservative and restrained revision, but neither moderate nor radical revision, satisfy Preservation restricted to flat conditionals, that is, FCP.[17]

Finally, we have to ask whether conservative and restrained revision validate Preservation for the full language. The answer is no. Consider again the blank slate $\mathcal{O}$ and two propositional atoms p and q. The

[14] This is the essence of a few key observations of Rott (2011, §8, on 'the subset problem').

[15] And, by the same token, Vacuity is trivially false.

[16] Note that $s(\mathcal{O}) \subseteq s(\mathcal{B}_2)$ only holds as long as no right-nesting is allowed. See below.

[17] That FCP is equivalent to structure preservation was observed independently by Leitgeb (2010, pp. 29–30) and Rott (2011, p. 176). Conservative and restrained revision also satisfy the slightly stronger condition of Vacuity restricted to flat conditionals.

 © Rott 2017

sentence $\neg p$ is consistent with $\mathcal{O}$. The conservative or restrained revision of $\mathcal{O}$ by $\neg p$ is the belief state $\mathcal{B}_2$ as defined above. Iterating either conservative or restrained revision, it is easy to check that $(\mathcal{O}*(p \vee q))*(p \vee \neg q) = \preceq_3$ with $v \preceq_3 w$ iff $v \notin [p \vee q]$ or $w \in [p]$ or $v, w \in [\neg p \wedge q]$, and $(\mathcal{B}_2*(p \vee q))*(p \vee \neg q) = \preceq_4$ with $v \preceq_4 w$ iff $v \in [p]$ or $w \notin [p \vee q]$ or $v, w \in [\neg p \wedge q]$. We find that p is true at all maximal worlds in the former belief state, while it is false at all maximal worlds in the latter belief state, so p is in $s((\mathcal{O}*(p \vee q))*(p \vee \neg q))$ but not in $s((\mathcal{B}_2*(p \vee q))*(p \vee \neg q))$. Using the Ramsey test based on either conservative or restrained revision, we find that $(p \vee q) > ((p \vee \neg q) > p))$ is supported by $\mathcal{O}$ but not by $\mathcal{B}_2$. Thus we have a violation of Preservation for right-nested conditionals.

The result of our semantic reflections on the situation for the full language is disenchanting: Inclusion does not even formulate a substantive requirement, and Preservation is not satisfied by any of the four canonical constructions![18] The legitimacy of these constructions does not depend on their satisfying any particular postulates; they are fully justified by semantic intuitions. I conclude that as long as one wants to keep the Ramsey test (which I took as a premiss for this discussion note), there is no point in aiming at the validation of Inclusion or Preservation. We are not facing the regrettable violation of two precious postulates, but Inclusion and Preservation are two conditions that simply cannot and should not be transferred from factual to conditional languages. It has been wrong from the very beginning to take it for granted that such a transfer makes sense.

A radically different interpretation of the above observations would be that all canonical constructions, or indeed the very representation of belief states in terms of plausibility preorders, have to be rejected. At present there is little evidence for this conclusion, but it may of course turn out to be right at some later time.[19] If it is, then alternative representations of belief and constructions for belief revision are called for.

Bracketing this option for the moment, how should we answer Rabinowicz's (1996) question 'Is Preservation worth preserving?' in the context of the full language? It has become clear that

[18] It is not difficult to verify that radical revision is the only canonical construction satisfying Bradley's (2007, p. 12; 2012, p. 152) principle of Commutativity, which is, given the Ramsey test, a weakened version of the Import-Export principle (applying only to 'indicative' or 'open' conditionals).

[19] Booth and Chandler (2016) offer an argument that may ultimately underpin this conclusion.

 © Rott 2017

Preservation cannot be preserved, and I do not know of any positive arguments that it ought to be preserved. The two canonical constructions that are favourable to the Import-Export principle for conditionals abandon Preservation even for flat conditionals. In so far as one considers Import-Export as essential for conditionals,[20] Preservation does not seem to be worth preserving at all for languages admitting Ramsey test conditionals. If this conclusion is correct, a major part of the new debate is deprived of its project.

5. Belief change and defeasible reasoning: defending Inclusion and Preservation of factual sentences

In his latest contribution to the new debate, Bradley hardly talks about conditionals and the Ramsey test. Instead, he claims that what we have learned from the debate is that the combination of a supra-classical consequence relation together with Inclusion and Vacuity gives a system of reasoning that 'does not leave enough space for defeasible reasoning' (Bradley 2017, p. 608). Bradley thus redirects our attention to the principal ideas underlying Inclusion and Vacuity, ideas that are quite independent of the problems of Ramsey test conditionals. Moreover, Bradley (2012, pp. 155–9; 2017) argues that relations of undermining and supporting should play an important role in belief change.[21] Again, this is a topic that does not appear to be intrinsically related to the Ramsey test analysis of conditionals.

Although Inclusion and Preservation look very persuasive *for a purely factual language*, it is not obvious that they ought to be satisfied. The failure of Inclusion is characteristic of 'inductively extended belief revision' as modelled by Levi (1996, pp. 142–51) and of 'abductive belief revision' as modelled by Pagnucco (1996) and Nayak (2000). Preservation fails to hold if the possible worlds are only *partially* ordered by $\preceq$ with respect to their plausibility.[22] Noteworthy arguments against Preservation were put forward in the context of Cross's

[20] Bradley and Chandler (see Chandler 2017, pp. 591–595) as well as Rott (2011, pp. 188–9) seem to sympathize with such a position, but none of them is very explicit about it.

[21] 'Supporting' now in the sense of 'backing', 'sustaining'. Chandler makes a similar point.

[22] Here is a very simple example. Again take atoms p and q, and let $\mathcal{B}_5 = \preceq_5$ be defined by $v \preceq_5 w$ iff either $v \notin [q]$ and $w \notin [p \wedge q]$, or $v \notin [p \wedge q]$ and $w \in [\neg p \wedge q]$, or $v, w \in [p \wedge q]$. Notice that $F = W$ and that $\preceq_5$ is reflexive and transitive, but *not* connected over F (to see this, take any $v \in [p \wedge q]$ and $w \notin [p \wedge q]$). Then $K = s^f(\mathcal{B}_5)$ is determined by $[q]$. So q is in K and p is consistent with K. But $K*p = s^f(\mathcal{B}*A)$ is determined by $\max_{\preceq_5}([p]) = [p]$. Thus q is not in $K*p$, contradicting Preservation.

 © Rott 2017

(1990) revision based on 'non-monotonic reasoning policies', Levi's (1996, pp. 29–42, 59–65) 'Ramsey revision', Rabinowicz's (1996) 'stable revision', Rott's (2001, ch. 5) 'foundational belief change', and Lin and Kelly's (2012) 'sensible, conditioning-tracking acceptance rules'.

Bradley's new twist is surprising in view of the fact that in a series of papers culminating in Gärdenfors and Makinson (1994), the authors showed that belief revision theory can serve as a theory of non-monotonic reasoning if it is applied to sets of 'expectations' rather than to sets of beliefs. In widespread terminology, *non-monotonic reasoning* is synonymous with *default reasoning* or *defeasible reasoning*. Gärdenfors (1990) went so far as to suggest that belief revision and non-monotonic logic are 'two sides of the same coin'. So how can there be a problem with defeasible reasoning?

Bradley suggests that Inclusion stands in the way of ampliative expansions and prevents well-supported beliefs from being accepted, while Preservation stands in the way of undermining and prevents poorly supported beliefs from being given up. Bradley gives an example that is meant to show that a belief B can be undermined by some information A that does not contradict the prior belief set.

> Suppose, for instance, that I measure the temperature of a glass of water using a thermometer and, as a result, come to believe that (B) it is 15 degrees Celsius, but that I am subsequently told that (A) the thermometer has not been functioning correctly for some time. I should be free to abandon my belief about the temperature of the water even if I did not explicitly hold the prior belief that the thermometer was functioning correctly. For I cannot reasonably be expected to anticipate all the things that I might learn that would undermine my current belief were I to learn that they were true. (Bradley 2017, p. 604; variables 'A' and 'B' inserted by me)[23]

It is doubtful whether this example achieves what Bradley intends it to achieve, because it fails to describe a credible scenario. Since B is a belief but $\neg A$ is not a belief, B is true at every world doxastically preferred by the believer, and there is some such world at which A is true. However, it seems incoherent of the believer to admit a world where the temperature of the water is 15 degrees (B) and the thermometer—which indicates 15 degrees!—is unreliable (A), and at the same time to exclude any world where the temperature of the water

[23] An earlier version of this example features in Bradley (2012, p. 155). Similar examples were given by Cross (1990, p. 232) and Rabinowicz (1996, pp. 105–6).

 © Rott 2017

is not 15 degrees and the thermometer is unreliable. If she reckons on the possibility of an unreliable thermometer, she should also reckon on the possibility of an incorrect display of the temperature. There is an ambiguity or vagueness in the belief predicate here. In a loose and everyday sense of the word 'believe', we are perfectly justified in saying that we believe that the temperature of the water is 15 degrees—and that the thermometer is working properly. Suppose, however, a stricter standard is applied, and more certainty is demanded for the (self-) ascription of belief, such that we can justly be said not to believe that the thermometer is working properly. Then we are losing the belief that the water is 15 degrees as well. In the loose understanding, the information that the thermometer has been unreliable is belief-contravening; in the strict understanding, the believer has never believed that the water has a temperature of 15 degrees Celsius. Either way, once the standard of strictness has been fixed, the situation described by Bradley cannot arise.

Having said that, I think that Bradley is right in pointing to the fact that there is no obvious way of placing defeasible reasoning within the theory of belief revision, or, what comes to much the same thing, of putting together beliefs and expectations in a single cognitive state and showing how such a state is or ought to be modified by new information. Thus we have learned from the new debate on the Ramsey test that this is an open problem that should be put on the research agenda.

6. Conclusion

I have argued that the new debate has relied too much on the method of postulation, and that it is more enlightening to bring together the discussion of the Ramsey test and the constructive models of iterated belief revision that were developed independently of this discussion in the 1990s. Focusing on four semantically motivated 'canonical constructions', I have advanced the thesis that the principles of Inclusion and Preservation do not have any normative justification in a language containing (possibly complex) conditionals. For the factual language, however, the attack on Inclusion and Preservation based on the idea of defeasibility has not succeeded. Though Bradley and Chandler are right in asking how underminers and supporters might be incorporated in belief revision theory, the real challenge for future research on defeasibility in belief change seems to lie in a

 © Rott 2017

systematic logical treatment of the simultaneous revision of beliefs and expectations.[24]

Appendix: The canonical constructions of belief state revisions

The four canonical constructions are different recipes for how to revise a belief state $\mathcal{B} = \preceq$ by a factual input A into a belief state $B * A = \preceq^*_A$. We present the basic definitions.[25] They are meant to apply to revisions of $\mathcal{B}$ by *doxastically possible* sentences, that is, sentences A such that $[A]$ has a non-empty intersection with the field $F_{\preceq}$ of $\preceq$. If this intersection is empty, $\mathcal{B}*A$ is always supposed to be the absurd belief state $\mathcal{A} = \varnothing$. The following definitions are supposed to be quantified over all v and w in $F_{\preceq}$.

Radical revision (also known as 'irrevocable' revision):

$$v \preceq^*_A w \text{ iff } \quad v \preceq w \text{ and } v, w \in [A]$$

Moderate revision (also known as 'lexicographic' revision):

$$v \preceq^*_A w \text{ iff } \begin{cases} v \preceq w \text{ and } (v \notin [A] \text{ or } w \in [A]) & \text{or} \\ v \notin [A] \text{ and } w \in [A] \end{cases}$$

Restrained revision:

$$v \preceq^*_A w \text{ iff } \begin{cases} v \prec w \text{ and } v \notin \max_{\preceq}([A]) & \text{or} \\ v \preceq w \text{ and } (v \notin [A] \text{ or } w \in [A]) & \text{or} \\ w \in \max_{\preceq}([A]) & \end{cases}.$$

[24] I am grateful to John Cantwell, Jake Chandler, Paul Égré, Tim Kraft, Arthur Paul Pedersen, Greg Wheeler, and an anonymous reviewer for valuable comments on an earlier version of this paper.

[25] For a short summary and diagrams, see Rott (2011, pp. 170–2). With the exception of restrained revision, the names of these constructions are mine. Here is a short list of the most relevant literature: for moderate revision, see Nayak (1994); for conservative revision, see Boutilier (1996); for radical revision, see Segerberg (1998); for restrained revision, see Booth and Meyer (2006). Except for radical revision, the canonical constructions satisfy the more general iteration postulates of Darwiche and Pearl (1997). Radical belief revision looks *very* radical, but perhaps it is just what we need for the evaluation of conditionals, because it satisfies the unrestricted Import-Export Principle $(K*A)*B = K*(A \wedge B)$ for factual A and B—even if A and B are logically incompatible with each other (thus contradicting the basic AGM postulates of Success and Consistency).

 © Rott 2017

Conservative revision (also known as 'natural' revision):

$$v \preceq^*_A w \text{ iff } \begin{cases} v \preceq w \text{ and } v \notin \max_{\preceq}([A]) \text{ or} \\ w \in \max_{\preceq}([A]) \end{cases}$$

It is easy to check that radical revision usually leads to a restriction of the field and that it may easily produce the absurd belief state $\mathcal{A}$, but that the field remains constant with the other constructions. A moderately revised plausibility relation $\preceq*$ is a refinement of a radically revised plausibility relation $\preceq*$, and a restrainedly revised plausibility relation $\preceq*$ is a refinement of a conservatively revised plausibility relation $\preceq*$.

References

Adams, Ernest W. 1970: 'Subjunctive and Indicative Conditionals'. *Foundations of Language*, 6, pp. 89–94.

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson 1985: 'On the Logic of Theory Change: Partial Meet Contraction and Revision Functions'. *Journal of Symbolic Logic*, 50, pp. 510–30.

Booth, Richard, and Jake Chandler 2016: 'The Irreducibility of Iterated to Single Revision'. *Journal of Philosophical Logic*, pp. 1–14. doi: 10.1007/s10992-016-9404-z.

Booth, Richard, and Thomas Meyer 2006: 'Admissible and Restrained Revision'. *Journal of Artificial Intelligence Research*, 26, pp. 127–51.

Boutilier, Craig 1994: 'Normative, Subjunctive and Autoepistemic Defaults' In Gerhard Lakemeyer and Bernhard Nebel (eds.), *Foundations of Knowledge Representation and Reasoning*, pp. 74–97. Berlin: Springer.

——1996: 'Iterated Revision and Minimal Change of Conditional Beliefs'. *Journal of Philosophical Logic*, 25, pp. 263–305.

Bradley, Richard 2007: 'A Defence of the Ramsey Test'. *Mind*, 116, pp. 1–21.

——2012: 'Restricting Preservation: A Response to Hill'. *Mind*, 121, pp. 147–59.

——2017: 'Supporters and Underminers: Reply to Chandler'. *Mind*, 126, pp. 603–8.

Chandler, Jake 2017: 'Preservation, Commutativity and *Modus Ponens*: Two Recent Triviality Results'. *Mind*, 126, pp. 579–602.

 © Rott 2017

Cross, Charles B. 1990: 'Belief Revision, Nonmonotonic Reasoning, and the Ramsey Test' In Henry E. Kyburg, Ronald P. Loui and Greg N. Carlson (eds.), *Knowledge Representation and Defeasible Reasoning*, pp. 223–44. Boston: Kluwer.

Darwiche, Adnan, and Judea Pearl 1997: 'On the Logic of Iterated Belief Revision'. *Artificial Intelligence*, 89, pp. 1–2.

Gärdenfors, Peter 1978: 'Conditionals and Changes of Belief' In Ilkka Niiniluoto and Raimo Tuomela (eds.), *The Logic and Epistemology of Scientific Change*, pp. 381–404. Amsterdam: North Holland.

——1986: 'Belief Revisions and the Ramsey Test for Conditionals'. *Philosophical Review*, 95, pp. 81–93.

——1988: *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Cambridge, MA: Bradford Books, MIT Press.

——1990: 'Belief Revision and Nonmonotonic Logic: Two Sides of the Same Coin?' In Luigia Carlucci Aiello (ed.), Proceedings of the Ninth European Conference on Artificial Intelligence (ECAI-90), pp. 768–73. London: Pitman.

Gärdenfors, Peter, and David Makinson 1994: 'Nonmonotonic Inference Based on Expectations'. *Artificial Intelligence*, 65, pp. 197–245.

Gärdenfors, Peter, and Hans Rott 1995: 'Belief Revision' In Dov M. Gabbay, Christopher J. Hogger, and John A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 4: *Epistemic and Temporal Reasoning*, pp. 35–132. Oxford: Oxford University Press.

Grove, Adam 1988: 'Two Modellings for Theory Change'. *Journal of Philosophical Logic*, 17, pp. 157–70.

Hansson, Sven Ove 1992: 'In Defense of the Ramsey Test'. *Journal of Philosophy*, 89, pp. 522–40.

Hill, Brian 2012: 'Defending the Ramsey Test: What is Wrong with Preservation?' *Mind*, 121, 131–46.

Katsuno, Hiromumi, and Alberto O. Mendelzon 1991: 'Propositional Knowledge Base Revision and Minimal Change'. *Artificial Intelligence*, 52, pp. 263–94.

Leitgeb, Hannes 2010: 'On the Ramsey Test without Triviality'. *Notre Dame Journal of Formal Logic*, 51, pp. 21–54.

Levi, Isaac 1996: *For the Sake of Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning*. Cambridge: Cambridge University Press.

 © Rott 2017

Lin, Hanti, and Kevin T. Kelly 2012: 'Propositional Reasoning that Tracks Probabilistic Reasoning'. *Journal of Philosophical Logic*, 41, pp. 957–81.

Lindström, Sten, and Włodzimierz Rabinowicz 1998: 'Conditionals and the Ramsey Test' In Dov M. Gabbay and Philippe Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 3: *Belief Change*, pp. 147–88. Dordrecht: Kluwer.

Makinson, David 1985: 'How to Give It Up: A Survey of Some Formal Aspects of the Logic of Theory Change'. *Synthese*, 62, pp. 347–63.

Nayak, Abhaya 1994: 'Iterated Belief Change Based on Epistemic Entrenchment'. *Erkenntnis*, 41, pp. 353–90.

——2000: 'Acceptance Without Minimality' In Manuel Ojeda-Aciego, Inman P. de Guzmán, Gerd Brewka, and Luís Moniz Pereira (eds.), *Logics in Artificial Intelligence. Proceedings of the European Workshop (JELIA 2000)*, pp. 163–78. Berlin: Springer.

Nute, Donald, and Charles B. Cross 2001: 'Conditional Logic' In Dov M. Gabbay Franz Guenthner (eds.), *Handbook of Philosophical Logic*, vol. 4, 2nd edn., pp. 1–98. Dordrecht: Kluwer.

Pagnucco, Maurice 1996: *The Role of Abductive Reasoning Within the Process of Belief Revision*. Unpublished doctoral dissertation, Basser Department of Computer Science, University of Sydney.

Rabinowicz, Włodzimierz 1996: 'Stable Revision, or Is Preservation Worth Preserving?' In André Fuhrmann and Hans Rott (eds.), *Logic, Action, and Information: Essays on Logic in Philosophy and Artificial Intelligence*, pp. 101–28. Berlin: de Gruyter.

Rott, Hans 1989: 'Conditionals and Theory Change: Revisions, Expansions, and Additions'. *Synthese*, 81 91–113.

——2001: *Change, Choice and Inference: A Study in Belief Revision and Nonmonotonic Reasoning*. Oxford: Oxford University Press.

——2011: 'Reapproaching Ramsey: Conditionals and Iterated Belief Change in the Spirit of AGM'. *Journal of Philosophical Logic*, 40, pp. 155–91.

Russell, Bertrand 1919: *Introduction to Mathematical Philosophy*. London: Allen and Unwin.

Segerberg, Krister 1998: 'Irrevocable Belief Revision in Dynamic Doxastic Logic'. *Notre Dame Journal of Formal Logic*, 39, pp. 287–306.

© Rott 2017

Book Reviews

The Logical Structure of Kinds, by Eric Funkhouser. Oxford: Oxford University Press, 2014. Pp. 182.

This brief but dense monograph has virtues whose impact it is not easy to assess. The book's central idea, both interesting and plausible, concerns certain key distinctions with respect to the relationships between kinds, properties, and concepts. Funkhouser aims to set out the methodological tools for an approach that he regards as promising (Funkhouser 2014, p. 137), instead of getting to the bottom of specific philosophical questions with the use of the tools: that's why the book's impact isn't easy to assess. Funkhouser defers to science to produce examples he's willing to wait for (Funkhouser 2014, p. 110), so the reader is put on hold. In what follows, I'll elaborate this aspect of Funkhouser's interesting work.

Funkhouser begins, helpfully, with a taxonomic look at relations of necessity between kinds. He goes on:

> The relations that are of particular interest to me are necessitation relations among kinds. My main goal in this book is to give an account of two such synchronic necessitation relations and illustrate their importance to scientific and metaphysical theorizing as well as the importance of our *recognizing the differences* between these relations. I will argue that understanding the differences between these necessitation relations is essential to adequately understanding and fruitfully employing scientific kind terms and the taxonomies to which they give rise. (Funkhouser 2014, p. 6, italics in original)

These are *determination* and *realization* relations. Believing that these have been conflated, Funkhouser articulates each clearly and plausibly, thereby distinguishing them. He suggests that this can enlighten the metaphysics of science, and more specifically, that the approach will lead to a better understanding of 'the structure of scientific concepts and the kinds they pick out' (Funkhouser 2014, p. 6). To be sure, the specific distinctions he articulates have application to other concepts besides scientific ones, such as the ethical concepts investigated since Plato, e.g., *courage* or *virtue*. The lessons are broadly applicable because 'I am interested in the logical form of these kinds rather than looking for particular essences' (Funkhouser 2014, p. 4). *Scarlet*'s relation to *red*, *water*'s relation to H_2O, *cup*'s relation

© Mind Association 2017

to *plastic-in-such-and-such-shape* and *courage*'s relation to *virtue* can all, Funkhouser argues, be enlightened by this central distinction.

Can his central distinction enlighten science and other fields? My assessment is a cautious yes. After all, progress even on what might appear initially to be side issues can pave the way for deep understanding of any area of inquiry. Confusions are like loose threads: they work their way to apparently remote material. Confusion about the contingent a priori can and arguably does lead to confusion and apparent paradox in unsuspected areas of philosophy: Kripke recalls that considerations about the contingent a priori, not a topic for which he is much celebrated, led in his mind to a battery of key clarifications concerning designation and necessity (Kripke 1980, pp. 14-5; a similar example is discussed in LaPorte 2003). Again, that is hardly surprising given the interconnectedness of different issues. On the other hand, sometimes Funkhouser appears, at least on first blush, to think or suggest that he's providing an instrument favouring some sort of interesting place for a non-functional, non-physical understanding of mental concepts (see e.g., Funkhouser 2014, pp. 61-2, 92, 116, 145-6; cf. Kripke: 'the analytical tools we are using go against the identity thesis and so go against the general thesis that mental states are just physical states' (Kripke 1971, p. 163; see also Kripke 1980, pp. 100, 155)). I don't think Funkhouser makes much progress here.

I will go through Funkhouser's main distinction and its application as he sees it; then I will discuss where the moral seems to reach its limit. As Funkhouser explains it,

> My main claim concerning the structure of kinds is that they have distinctive dimensions that partially capture their essences, and these dimensions determine the determinates of those kinds while also distinguishing them from their realizations. (Funkhouser 2014 p. 9)

Scarlet and especially some maximally specific shade of scarlet relate differently to *red* than any neural state relates to a kind of pain. To say otherwise is a confusion that allegedly infects the work of important contributors to philosophical psychology; anyway, it's a natural confusion to make. 'Why not say that NS_1 is a kind of *pain*, just like *MS-shade*$_1$ is a kind of *red*? I imagine that this will be a common reaction' (Funkhouser 2014, p. 22, italics in original). The main problem, as Funkhouser sees it, is that 'two or more maximally specific neuroscientific kinds can, as far as we know, kind-necessitate the same maximally specific *pain* kind' (Funkhouser 2014, p. 22, italics in original). NS_1 exemplified in some humans and NS_2 exemplified in other humans or indeed other species, can be identical so far as their painfulness is concerned—e.g., they can have the same place on the sharp–dull spectrum and intense–weak spectrum, say, giving the specific feel a point in a chart of the logical space of *pain*. Compare two maximally specific shades of *red*: *scarlet*$_1$ and *brick-red*$_{452}$. These will vary, according to the relevant determination dimensions (say, the hue, brightness, and saturation spectra) in such a way that they simply couldn't be identical in their *red*ness.

 © Mind Association 2017

The basic idea here is plausible and beautiful. To set before you another example, consider the rectangularity of a table top. Clearly that's different from its woodenness. Rectangularity might need instantiation in some material—plastic or glass or wood—but instantiated rectangularities can be *identical* as to their *rectangularity* so that two table tops of different material can both have the *same* maximally specific rectangularity, as determined by side number and lengths. This, then, is multiple realization. Now consider determination: if the dimensions of determination differ (length of sides, say) then you have different rectangles—their rectangularity as such is different.

What is saliently unclear or doubtful to me is whether this distinction is a game changer in, say, philosophical psychology, which is Funkhouser's main area of application for demonstrating the significance of the distinction. Recall that *pain*, when multiply realized, is the same sensation in different realizations: so, in an interesting respect, its realizer is irrelevant. An expert in cognitive science could study *pain* without understanding the specifics of neuroscience and hence without understanding much about realizers of *pain*, whether neuroscientific, silicone-based, or whatnot (Funkhouser 2014, p. 113). There is a kind of explanatory autonomy, as well as an ontological autonomy. Funkhouser expresses confidence or perhaps optimism that scientific development will confirm examples like this 'in which special science generalizations hold across objects or systems that differ wildly in their realizations' (Funkhouser 2014, p. 108).

Examples to compare to philosophical psychology could in my view be better developed. But take an example that Funkhouser borrows from Fodor, quoting from the latter: 'Does anybody really think that … mountains-as-such won't continue to serve geology in good stead', even though mountains can be made of different materials (Funkhouser 2014, p. 110)? Let us appeal to this example. Clearly one can maintain autonomy of the special sciences without addressing what has been in my view the main question about pain in the literature from Kripke, in familiar traditions, and indeed well before Kripke, in anticipatory traditions, namely, whether having made any neurophysiological or other realization of pain, like a c-fibre firing, God would have any *further* work to do in the world to make the c-fibre firing painful, or to make *pain*. Even if *mountainhood* or *mountain-of-granite*-hood (the refinement simplifies the presentation) enjoys ontological status *distinct* from that of kinds like *tall hunk of 'light-colored igneous rock with grains large enough to be visible with the unaided eye'* (King 2005-2016; feel free to reformulate this characterization of *mountain-of-granite*-hood), it isn't as if, after God put together molecules of a certain kind in a certain formation—a tall hunk of light-coloured igneous rock with grains large enough to be visible with the unaided eye—any further work would have to be done to make a mountain of granite. Or think of *lubricants* realized in microstructural sheets like those forming graphite, but also realized by cardboard sheets stacked to slide (my example, not Funkhouser's). Having made the sheets with all their

 © Mind Association 2017

specifications, it would seem settled that we have a lubricant. Of course there are details to be worked out here—e.g., concerning ceteris paribus qualifications, the dispositional nature of a lubricant, etc. But there is nothing in all this that suggests God is going to have to do more to create a lubricant than to make sheets held together by strong bonds within sheets but weak bonds uniting one sheet with another in a vertical stack, so that the sheets easily slide apart from one another. C-fibre firing seems different: again, God seems to have more to do in order to make pain than to make c-fibre firings.

The standard physicalist positions on pain suggest ready adaption to Funkhouser's logic of kinds; not much reformulation seems needed. So take 'new wavers' who argue that physical kinds are *ontologically* identical to phenomenal ones despite a *conceptual* discontinuity. They could argue, in Funkhouser's framework, that ontologically, pain's intensity-dimension is mirrored in some span of intensity in neurophysiological kinds, so that there's not really multiple realizability but rather determination: for any neuro-kind, there'll be a somewhat distinct phenomenal feel, however slight, whether perceived or not. (Others besides new wavers will take something like this response, as Funkhouser sees: e.g., David Lewis, addressed at Funkhouser 2014, p. 148). New wavers' main pillars remain in place: they still maintain that neuro-kinds are identical to phenomenal kinds and that even so, there's conceptual distinctness, so that we can't learn a priori of the identity although there's an a posteriori reason to affirm the identity. And this response to Funkhouser's categories belies outright his hunch that new wavers 'conceive of themselves as objecting, in some sense, to arguments like our MR argument for Ontological Autonomy' (Funkhouser 2014, p. 149).

Other possible reformulations on the part of new wavers would force them to recognize Funkhouser's ontological distinctness between the physical and phenomenal—and so would force them to qualify any crude statement of their position of ontological monism—but even so, will allow them to keep the spirit of their views. For example, new wavers might alternatively concede that the 'individuation conditions' (Funkhouser 2014, p. 92) of pains differ from those of neuro-kinds. That concession would allow Funkhouser to drive his ontological wedge but the spirit of new wavers' claims to remain intact despite Funkhouser's ontological dualism because new wavers could go on to clarify that nomologically and metaphysically, a full account of a neuro-kind articulates its painfulness in the same way a full account of cardboard stacks, with their bonds, tighter and looser, articulates how they can serve as a sliding lubricant for moving a piano; you don't have to account for the sliding and then go on to add that you have a lubricant. The key difference, according to new wavers, between the case of lubricants and the case of pain, is that an account of c-fibre firing, despite its scientific and metaphysical adequacy as the whole story, yields only a weak or yet-to-be-provided a posteriori understanding of why the pain arises.

 © Mind Association 2017

Given these and other, messier options, the moral of Funkhouser's framework becomes complicated, ramifying into different possibilities, but in any case it need not seem problematic to new wavers. Other materialist positions like the two main varieties of functionalism also seem to me to accommodate Funkhouser's framework without much trouble. So it isn't clear to me that we have a game changer in philosophical psychology. Funkhouser draws an interesting distinction, and it's very hard to believe that it can't be used to clear away some of the confusions at the root of philosophical and scientific puzzles. So, for example, the distinction seems important for understanding how it is that *orange* is between *red* and *yellow*, so comparable to these colors, but that *orange* is relatively incomparable with the taste *sour* (Funkhouser 2014, pp. 53-4). Even so, it isn't yet easy to distil a ready moral or set of morals from the book's main propositions. Already an interesting read as an abstract study, the book may in time have a clear or unified philosophical pay out.

References

King, Hobart 2005-2016, 'Granite', Geology.com. Accessed July 5, 2016.
Kripke, Saul 1971, 'Identity and Necessity', in M.K. Munitz (ed.), *Identity and Individuation* (New York: New York University Press).
——1980, *Naming and Necessity* (Cambridge, MA: Harvard University Press).
LaPorte, Joseph 2003, 'Does a Type Specimen Necessarily or Contingently Belong to Its Species?', *Biology and Philosophy*.

Hope College JOSEPH LAPORTE
jlaporte@hope.edu

doi:10.1093/mind/fzw062 Advance Access publication 1 February 2017

The Possibility of Inquiry: Meno's Paradox from Socrates to Sextus, by Gail Fine. Oxford: Oxford University Press, 2014. Pp. xiv + 399.

Gail Fine's excellent book is a rich and sophisticated examination of an ancient epistemological puzzle first stated in Plato's *Meno*. After Socrates and Meno try unsuccessfully to discover what virtue is, Meno, frustrated by Socrates' questioning, accuses him of being a numbing torpedo-fish and is ready to quit (79e7–80b7). Socrates replies that he makes people numb with perplexity because he is in the same condition (80b8–d3). He then urges Meno to continue searching with him for what virtue is (80d3–4). Meno then says (all translations are Fine's, with her additions ('(M1)' etc.)):

> (M1) But how will you inquire into this, Socrates, when you don't at all know what it is? (M2) For what sort of thing, from among those you don't know, will you put

 © Mind Association 2017

> forward as the thing you're inquiring into? (M3) And even if you really encounter it, how will you know that this is the thing you didn't know? (80d5–8)

Socrates responds by recasting Meno's questions in the form of a dilemma, which Fine reconstructs as follows (p. 8):

> (S1) For any x, one either knows, or does not know, x.
> (S2) If one knows x, one cannot inquire into x.
> (S3) If one does not know x, one cannot inquire into x.
> (S4) Therefore, for any x, one cannot inquire into x.

'Meno's Paradox' is the conjunction of Meno's three questions and Socrates' dilemma.

In Part I, Fine discusses, with the rigour and nuance characteristic of her other work, the origins of the puzzle in the *Meno*, the different ways in which it can be read, and Plato's three-stage response. In Part II, she examines discussions of and responses to the puzzle in Aristotle, the Epicureans, the Stoics, Plutarch, and Sextus Empiricus. In what follows, I will focus on Part I, which can be read as a stand-alone, book-length treatment of the epistemology of the *Meno*. Here Fine amplifies and defends the interpretation of the *Meno* for which she is justly famous. It thus marks a very significant contribution to the literature.

Meno's three questions challenge the possibility of beginning and successfully concluding an inquiry. (M1) and (M2) raise what Fine calls 'the Targeting Objection': 'one can't inquire into something if one doesn't at all know what it is since, in that case, one won't have a target to aim at' (p. 75). (M3) raises what she calls 'the Recognition Objection': 'if one doesn't antecedently know the thing one is inquiring into, one won't be able to know, or realize, one has found it if one manages to do so' (p. 80). Fine points out that (among other differences) Socrates' dilemma drops the Recognition Objection and adds the worry about seeking that which one already knows (S2) (pp. 84–5). Nonetheless, she argues persuasively that Meno presents a single, if complex, puzzle, which Socrates then extends while still capturing his concerns (p. 104). For Meno's main worry, which motivates both objections, is that Socrates is in a 'cognitive blank' about virtue and no one in that condition can inquire into or discover what virtue is, and (S3) captures that concern.

Fine's interpretation of Plato's response to Meno's Paradox (Chapters 4–5) is the most interesting, and is likely to be the most controversial, part of her book. In Chapter 4, she argues that Socrates argues that (S3) is false. While it is true that one cannot inquire into something if one is in a cognitive blank about it, and while it is true that Socrates claims that he does 'not know at all what virtue is' (71a5–7), Meno is wrong to infer from this that Socrates is in a cognitive blank about virtue. For Socrates only disavows knowledge as he defines it later in the dialogue (98a): 'a true belief that is tied down with reasoning about the explanation' (p. 16). Fine calls this P-knowledge and

 © Mind Association 2017

points out correctly that Socrates' disavowal of it does not mean that he is in a cognitive blank. For example, one can lack P-knowledge and still have true beliefs. Since having relevant true beliefs is sufficient for inquiry, (S_3) is false. And since Meno and Socrates lack P-knowledge but have true beliefs about virtue, Meno's questions are answered: they can inquire into what virtue is and realize they have found it, should they manage to do so. Fine argues that Socrates attempts to prove all of this in the famous episode in which he asks one of Meno's slaves questions about a geometrical problem the solution to which, Socrates argues, the slave works out for himself.

In Chapters 4–5, Fine examines Plato's response to Meno's Paradox in the theory of recollection, statements of which bookend Socrates' examination of the slave (hence Plato's 'three-stage' response). Fine argues that the examination of the slave shows *that* inquiry is possible and the theory of recollection explains *how* (pp. 150–1). Inquiry is possible because we have relevant true beliefs we tend to rely on when we search. But what explains our tendency, vividly illustrated in the examination of the slave, to favour true beliefs over false? Plato's answer is recollection: the slave is remembering what he once knew. Most scholars think that recollection consists in remembering what we *once* knew prenatally and *still* know latently and innately. Fine disagrees. She argues that Plato is only a prenatalist and not an innatist about knowledge: inquiry and learning consist in recollecting not latent innate knowledge but knowledge we once had but have since completely forgotten.

Some scholars think that Plato commits himself to innatism in the following lines, in which Socrates reflects back on his examination of the slave:

> Won't he know without having been taught by anyone, but <only by> being questioned, taking up (*analabôn*) the knowledge himself from himself (*autos ex hautou*)? Yes. Isn't it the case that to take up (*analambanein*) knowledge oneself in oneself (*auton en hautô(i)*) is to recollect (*anamimnêskesthai*)? Certainly. (85d3–7)

Fine argues that the verb *analambanein*, which can mean 'to recover', 'can also mean … "to take up", in a sense that is compatible with … learning for the first time' (p. 150). However, as Whitney Schwab has pointed out (in his review of Fine's book for *Notre Dame Philosophical Reviews*), the troubling claim for Fine's interpretation is not that the slave is said to *analambanein* knowledge but that he is said to *analambanein* knowledge 'himself from himself' and that to *analambanein* knowledge 'oneself in oneself' is recollection. These phrases strongly suggest that the slave is 'taking up' knowledge that is already in him, and hence they strongly suggest a commitment to innatism.

Fine objects to the innatist interpretation of 85d3–7 by saying that if 'Socrates means that the slave is recovering knowledge he already has or once had, the remark … is wholly unjustified' (p. 150). However, Socrates seems to think that the remark is justified by the preceding discussion: he is saying that the best, and perhaps the only, explanation of the slave's ability

 © Mind Association 2017

to work out the solution to the geometrical puzzle for himself is that knowledge of it is already in him.

Here is an additional worry about Fine's interpretation. At the conclusion of his discussion of recollection, Socrates says: 'Then if (i) the truth about beings is always in the soul, (ii) won't the soul be immortal, so that (iii) you must confidently try to inquire into, and recollect, what you don't happen to know now: that is, what you don't remember?' (86b1–4; I have added '(i)' etc.). My worry is that it is hard to see how (i), read in Fine's way as a statement only of prenatalism, provides any support for (iii). Why should the fact that Meno once had knowledge, but now no longer has it at all, give him any reason to feel confident that he can inquire? Fine's answer seems to be that since Meno's soul once knew, he is now disposed to favour true beliefs over false. However, Socrates never says this, and it's hard to imagine Meno, no paragon of philosophical acumen, inferring that this is what Socrates means (or means to suggest) by (i). If we read (i) as a statement of innatism, then it seems clearer why Meno should feel confident that he can inquire: the knowledge he'd like to possess is actually already in him. Evidently, Socrates introduces the theory of recollection in part to encourage Meno to inquire despite his confusion. The innatist, more so than Fine's prenatalist, interpretation of recollection is in a good position to make sense of this.

In Part 2, Fine argues that in their accounts of inquiry 'Plato, Aristotle, the Epicureans, and the Stoics are closer to one another than they are sometimes taken to be, and in surprising ways' (p. 299). For example, they all agree that P-knowledge in this life is not required for inquiry and they all reject the existence of innate knowledge. Nonetheless, in Plutarch she finds, and examines the merits of, an argument that among Plato, the Peripatetics, the Epicureans, and the Stoics, 'only Plato has a satisfactory solution to Meno's Paradox' (p. 318). Fine concludes (Chapters 10–11) by examining two challenges, one each from the Stoics and the Epicureans, to the possibility of Skeptical inquiry and Sextus' responses. Although among the ancient authors Fine discusses in Part 2, only Aristotle and Plutarch mention Meno's Paradox explicitly, she argues that post-Platonic accounts of inquiry are motivated by many of the concerns introduced in the *Meno*.

This book is an example of philosophically informed scholarship at its very best. It is essential reading for anyone with interests in any aspect of ancient epistemology.

Georgetown University DAVID BRONSTEIN
david.bronstein@georgetown.edu
doi:10.1093/mind/fzw071 Advance Access publication 10 January 2017

 © Mind Association 2017

Aesthetics as Philosophy of Perception, by Bence Nanay. Oxford: Oxford University Press, 2015. Pp. 240.

Imagine you are looking at a landscape. The dusk, calm and long, spreads across the grassy plain, the setting sun sheds strains of light beneath an old bridge, covering the land with blurry golden stripes. Whether this landscape is real or depicted on a canvas, the experience seems pretty special. Classically, we would call this an *aesthetic* experience, by contrast with the more mundane experiences you could have had if you simply kept walking, ignoring the evening shades altogether, or if you had looked at the bridge and wondered when it was built. As you stop and stare at the landscape, the experience seems not to be special in terms of the objects you perceive, but in the way you are 'taking in' the scene and its richness.

Defining what unifies and distinguishes these special, aesthetic cases from non-aesthetic ones constitutes one of the key problems of philosophical aesthetics. Nanay's book, as its title suggests, sees the solution as lying with theories of perception — especially theories of attention, cognitive penetration, and mental imagery. Add to this that theories of perception, as Nanay conceives them, need to be connected to cognitive neuroscience, and you can see that the broader ambition of the book is to help make the problem of the aesthetic difference more tractable in experimental ways.

Of course, there will be the easy objection that aesthetics should stay away from the experimental sciences. Or that a stable difference between aesthetic and non-aesthetic cases across contexts and objects is in itself questionable. But shall we for that reason simply go back to calling it a 'je ne sais quoi', or claim, as Noe (2011) does for art, that 'the question must arise, but it allows no definitive answer'?

Nanay's book, *Aesthetics as Philosophy of Perception*, does not engage with the critical discussions, past and present, about the status of experimental or neuro-aesthetics. (See Maynard (2016) for an example.) I think this is one of the main virtues of the book: showing what can be done if one goes down the road of linking aesthetics and empirically informed theories of perception, rather than commenting endlessly about how careful one needs to be before saying something general about aesthetics. The book is divided into eight chapters: Chapters 1 to 3 explain which part of philosophy of perception (i.e. mostly attention) is most likely to shed light on the nature of aesthetic experiences, while Chapters 4 and 5 examine what this changes for classical debates about aesthetic properties, or depiction. Chapters 6 and 7 stand more by themselves, with the first offering an account of aesthetic uniqueness, and the last a survey of aesthetic experiences other than the visual cases which are central to the previous discussions. There is a certain boldness in the tone and content of the book, but so there is in a sketch: the bravery needed to start outlining what a full picture could be like.

 © Mind Association 2017

The boldest proposal of the book is also the most interesting: that a distinctive kind of attention underlies the difference between aesthetic and non-aesthetic experiences, if not in all, then at least in most cases. To go back to the example above, what is going on in your mind when you seem to have a heightened or special experience of the landscape? The answer is considered to go along the following lines:

(i) You attend to the object (the landscape, the painting) in a specific *way, i.e. by distributing your attention over the properties of the object.*
(ii) Doing so, you gain access to *aesthetically relevant properties.*
(iii) This makes a difference to your experience (or, as we discussed before, to how much you enjoy your experience).

To be clear, this is presented as a research programme, rather than a fully-fledged account. As Nanay notes, generalisations in the domain of aesthetics, especially when dealing with art works, are controversial, and he welcomes exceptions. The ambition of the book, to ground the specificity of aesthetic experiences in differences in attention, is immediately presented as imperfect. Some aesthetic experiences may not require attention at all (p. 28); novels and narratives might require a different framework, where distributed attention is not central (p. 160 sq.). The role of attention, then, is not offered as a necessary condition for an experience to be aesthetic: it is not necessary for someone to distribute her attention over the properties of the landscape for her to gain access to aesthetically relevant properties. Having access to such properties, in one way or another, might still be necessary for her experience of the landscape to be aesthetic, which is a part of the discussion about aesthetic properties which I am not going to review here. What seems to me to be more interesting is whether a distinctive kind of attention will be sufficient to provide this access. Won't judgement or emotional responses need to play a role in many cases of aesthetic experiences? How does this distinctive kind of attention operate?

Perhaps these questions are going beyond what the book intends to do. After all, all we are told is that 'attending to a variety of properties of the object we are looking at — that is, aesthetic attention — is a central feature of some paradigmatic cases of aesthetic experience'. As we give up on finding necessary conditions for turning an experience into an aesthetic one, do we have to give up on the hope of finding sufficient ones? It is sometimes difficult to decide what Nanay thinks here: on the one hand, he seems to suggest that empirical generalisations, connected to conceptual distinctions, are the best one can hope for in aesthetics; on the other hand, he suggests that the book is a research programme, with testable predictions. I will follow the second suggestion here, and, in the rest of the review, focus on what I take to be a core proposal of the book: if we pick a paradigmatic case like viewing a real or painted landscape, a certain form of attention is sufficient to explain

 © Mind Association 2017

why our take on that landscape should count as aesthetic. More specifically, differences in how we attend to the landscape between the case in which we stop and stare and the case in which we simply walk through it will mean that our conscious perception of the landscape will be different in the two cases and only the first one will qualify as aesthetic. A difference in the way we distribute our attention over the scene turns our perceptual experience into an aesthetic experience.

1. Is the aesthetic difference really manifest in experience?

Nanay notes here that it is not 'a particularly novel claim to consider aesthetics to be about experiences' (p. 9). Still, what the word 'experience' includes and what it doesn't needs to be made clear, if this is what the aesthetic distinction (partly) rests on. As Paul Snowdon nicely reminds us in his (2010) paper, the boundaries of an experience are not always very well defined.

By accepting that the difference between the aesthetic contemplation of the landscape and our mundane seeing of the same landscape is a matter of experience, we are accepting that it is not *just* a matter of how the same experience leads to different judgements or emotions. Note that this is not to deny that there could be differences in judgements and emotional responses between the two cases: you might judge the landscape to be harmonious in one case and not in the other; you might feel joyful and rested, or remain neutral. But what the book asks us to consider is that these differences derive from, or come after, a more fundamental difference in the way you experience the landscape.

The title of the book makes us further believe that by 'experience' we should understand 'perceptual experience'. We see the landscape differently when we contemplate it and when we simply stare at it while we rest in the middle of our walk. The difference here should not be thought of in terms of how long the experience lasts, of course, nor in terms of what the broader object of experience is, i.e. the landscape. The difference we are invited to consider is that, when contemplating the landscape, we experience certain properties, especially relational or formal properties between objects and/or sensory properties, which are missed when we simply stare at it. What we are asked to think here is that the same object is experienced, but some properties, like balance, harmony, symmetry etc., are made salient in one case and not the other — and this explains why our experience qualifies as an aesthetic one.

This proposal is difficult to reconcile with another suggestion (which is also interesting in itself), i.e. that the aesthetic distinction could be not in how we experience the landscape, but in how much we *appreciate* our experience of that landscape (p. 73). Here the difference seems to be in the pleasure, or the emotions, derived from the perceptual experience, and not so much in a difference in the perceptual experience *per se*. After all, the idea

 © Mind Association 2017

that the difference between aesthetic and non-aesthetic cases has to do with enjoyment or pleasure, and not with perception, has both a long pedigree and an intuitive appeal.

What, however, is the problem with extending the word 'experience' to include enjoyment, pleasure, and emotions, and preserving the claim that aesthetics is 'about experience'? The difficulty here is to understand what attention has to do with how much we enjoy our experience. Could it be the case that distributing one's attention over the properties of a single object affects our appreciation of our experience? Why would that be more enjoyable than distributing attention over multiple objects for instance, as when we are looking for our dog in the middle of the grassy field? If the enjoyment is derived from the picking of certain relevant properties (say balance or harmony), and if it is this enjoyment that brings the distinctively aesthetic mark to our experience, then it is less clear why attention is so central, unless attention is necessary to pick up these properties (which we are told it is not). Alternatively, I wonder whether, if the distinctive aesthetic character of the contemplation of the landscape comes from a certain kind of enjoyment which is linked to attention, the book should not have looked into ways in which this enjoyment comes from our lack of immediate interest in detecting or identifying certain pre-determined objects or properties. Still, the more fundamental thing to explain might then be disinterest, rather than attention — and a lot has been written here on disinterested attitudes in aesthetics. The attentional research programme here still needs to stress why it does something different or better than these traditional accounts framed in terms of attitudes such as the ones offered by Bullough (1912) and reviewed in King (2016). An explanation in terms of attention would be easier to map onto experimental data, but more needs to be said to explain how attention and appreciation can be articulated with respect to one another.

This brings us to an even more difficult point. In Chapters 4 and 6, we read that distributing our attention over the properties of an object (i.e. 'aesthetic attention') will sometimes lead to a difference in judgement. For instance, when you contemplate the landscape, your attention makes you attend to its harmony or peaceful character. But harmony or peacefulness might be something you judge, not something you experience. You might also become sensitive to the similarity between this landscape and one of these ancient Hiroshige prints you saw at the British museum. Again, this similarity might be something constructed through judgement, and not given in experience.

Note that nothing here hinges especially on these examples: some philosophers will be happy to consider that harmony, peacefulness, or similarity relations are perceived (e.g Peacocke (1987) or Hopkins (1998) on experienced resemblance). More austere and more empirical accounts of perception will resist this. The important point is not to argue about specific properties. There will still be many cases where the properties which are distinctively

 © Mind Association 2017

accessed in the aesthetic cases are properties that need to be thought of and cannot be perceived. In simpler words: to a large extent aesthetics is about how we evaluate and judge what we see, and how we discuss this with others — and not so much about what we see. What, then, is the relevance of a research programme concerned with attention and perception, if aesthetics is about judgement?

There are two ways to respond here. One would be to accept that aesthetics is only partly about experience, and that philosophical insights about perception can then only illuminate cases which come with differences in experience. The other, which is favoured by Nanay, comes to the rescue of the primacy of experience by suggesting that what we think about the object of our aesthetic contemplation 'feeds back' into our on-going experience of that object. For instance, our judgement that the landscape is harmonious or looks like a Hiroshige print will change how we see it. Bracketing the issue about mechanisms, appealing to cognitive penetration of perception in the circumstances might look strategic, but is eventually costly: should the defence of a research programme in aesthetics commit us to a controversial claim about perceptual processing? Don't new models of the mind suggest more interesting ways to think about the interplay of higher-level predictions and incoming sensory signals as shaping our experience and interpretation of a scene?

The commitment to cognitive penetration also ends up being puzzling when brought in line with the core claim of the book, the claim that the distinctiveness of aesthetic experience derives from attention. Differences in attention are among the first things to be ruled out as establishing a case of cognitive penetration of perception (Pylyshyn 1999, Macpherson 2012, Deroy 2013). If there is a difference between how you attend to the drawing of the duck-rabbit which makes you experience it in one case as a duck, in the other as a rabbit, then this is all you need to explain a difference in experience. The same thing goes for the landscape: if there is a difference in how you attend to the scene between the contemplative and the non-contemplative case, then this might explain why you come to be sensitive to its harmony in one case, and not the other. Cognitive penetration and differences in judgements are not necessary to explain the difference in experience because attention does the job.

Perhaps one should not be overly concerned with the issue of cognitive penetration. I still think that this is an example of the real challenges that face the programme of linking aesthetics and philosophy of perception: is the idea that our theories of aesthetics require us to sign up for certain theories of perception? Shouldn't what we say about aesthetics be compatible with all theories of perception? What Nanay suggests — and what I am sympathetic to — is that aesthetics is actually responsive to how perception works (at least in large part). But this comes with challenges. The challenges become very visible when turning to the core claim of the book, which is that a

 © Mind Association 2017

distinctive kind of attention needs to be posited to explain aesthetic experiences.

2. Is there such a thing as aesthetic attention?

The idea that there is a link between aesthetic, advanced in the book, is not novel, but the book must be credited for bringing it within the spotlight, and making it central to questions about the role of attention in aesthetic experiences.

Saul Bellow, among others, remarked that 'art has something to do with an arrest of attention in the midst of distraction' (Bellow, 1966; p. 54). More than a difference in what we see, aesthetic contemplation seems to have to do with how we look. 'Arrest' seems a good way of capturing it. An austere account — by which I am somewhat tempted — would reduce aesthetics to the lack (or inhibition) of distraction, which extends the time we can spend focusing on something. Museums and galleries should be places of minimal distraction, and the stillness we adopt in theatres or concert halls is conducive to a more focused attention.

This is not what Nanay suggests however. According to him, aesthetic cases are not characterised by a difference in intensity or span of attention, but by a distinctive *kind* of attention that is both focused on an object (or scene) and distributed over its properties. This form of attention is said to differ from the kind of distributed attention already singled out in cognitive neuroscience, which is distributed over a region of space and samples properties and objects across this region. But why posit a new kind of attention to be at play in aesthetic cases, and do we have evidence that it exists?

It is fair to observe that Nanay's motivation to posit a distinctive kind of attention to explain aesthetic cases is not primarily driven by empirical results, as is the case for other kinds of attention discussed in the literature. Distributed attention (which covers a region of space) is shown to differ from focused attention (which selects one object) both behaviourally and functionally. Individuals can either choose to exert one or the other (in a top down, voluntary fashion) or be driven by circumstances to rely on one or the other (in a bottom up, involuntary fashion). Tracking of moving targets for instance requires focused attention and can only be performed successfully with up to four targets (e.g. Cavanagh & Alvarez 2005). Attention is involved both in selecting the relevant targets and in maintaining object information in short-term memory as the targets move. By contrast, other tasks like reporting statistical information across a set of objects are not subject to such limitations: participants who were poor at reporting the size of individual circles can accurately report the mean size of more than four circles (Ariely 2001). Performance in the mean size judgment task is little, if at all, affected by variation in the number or density of items (Chong & Treisman 2005). Distributed attention, then, is especially relevant for statistical estimation,

 © Mind Association 2017

and might be drawing on less structured working memory representations than in the focused cases (Baijal et al. 2013).

The notion of aesthetic attention, then, defined by Nanay as a hybrid between focus (on one object) and distribution (over properties), goes beyond this established distinction. Because it still draws on this distinction, however, it can be hard to square with it: if focused attention is, for instance, strongly linked to binding of features, what about the hybrid, semi-focused, semi-distributed attention? Are the features of the abstract or figurative painting, or of the landscape, bound, semi-bound, or unbound?

Fitting this new kind of attention with existing scientific evidence is, then, a major challenge. Part of the experimental credentials of the book comes from a study conducted by Vogt and Magnusen (2007), which examined differences between artistically-trained viewers and novices looking freely at pictures. Although there was no significant difference in the number or duration of fixation, the trained viewers tended to distribute their attention over a larger area, and attend more to structural features than specific objects. A different pattern was observed when the participants were instructed that the viewing of the picture would be followed by a memory test afterwards, suggesting that the distribution of attention was indeed linked to the adoption of an 'aesthetic attitude' by the experts. That the distribution of attention was also focusing on an object, however, is not evidenced by this protocol.

Turning to other studies, however, one might wonder, if the distribution of attention noticed in (expert) aesthetic viewing requires us to posit a distinctive kind of attention. According to Lochler et al. (2012), distributing attention (in the classical way) is what everyone does when viewing a work of art (no matter if it is abstract or figurative): all subjects sweep the image to get its 'gist' for about two seconds, before focusing on finer details. Still, in this case, what happens is not that a hybrid between distributed and focused attention is at play, but that one uses distributed attention then focused attention. The proposal that classical forms of attention operate serially is different from Nanay's, though seems to explain how we pick up on aesthetically relevant properties equally well: the initial distribution of attention helps the viewer form a hypothesis about the scene or object, which is then used to focus attention on certain properties. This proposal is more economical, as it only rests on two well-established forms of attention, and does not require us to posit a distinctive kind of attention.

What the book needs to establish more strongly is that neither focused nor distributed attention as they have just been described is suited to explain aesthetic cases. But why accept Nanay's hybrid theory, rather than the one I just outlined, i.e. the serial theory? The book certainly leaves the reader curious to hear more from Nanay. One argument that can be reconstructed from the book is that both focused and distributed attention often correspond to a strategic allocation of attention — one where the selection is done in a top-down way on a certain identified object ('detect the green tree',

 © Mind Association 2017

'track the four birds') or governed by a pre-determined task ('estimate the average size of the circles'). These two top-down forms of attention might not present the kind of freedom or disinterest that we expect from aesthetic cases. Is the hybrid form of attention posited for aesthetic experience different? Is there a way in which, having voluntarily decided to focus on a certain object, our attention is somewhat captured and guided across its properties, in a way that does not require so much top-down influence?

I have only discussed here two of the core themes of Nanay's research programme — the experiential and the attentional one, so to speak. Its main merit is certainly to point out ways in which they can be articulated with respect to one another, and to suggest new ways in which they could be jointly investigated. I look forward to seeing the future developments of this experiential-attentional research programme, both philosophically and experimentally. And I hope that it will help us understand what is so special about looking at the dusky, grassy plain, and its blurry golden stripes.*

Institute of Philosophy, School of Advanced Study, OPHELIA DEROY
University of London
Ludwig Maximilian University, Munich
ophelia.deroy@sas.ac.uk

References

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162.

Baijal, S., C. Nakatani, C. van Leeuwen, and N. Srinivasan (2013). Processing statistics: An examination of focused and distributed attention using event related potentials. *Vision Research*, 85, 20–25.

Bellow, S. (1996). The art of fiction. *The Paris Review*, 36, 48–73.

Bullough, E. (1912). 'Psychical distance' as a factor in art and an aesthetic principle. *British Journal of Psychology*, 5, 87–118.

Cavanagh, P. and G. A. Alvarez (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, 9(7), 349–354.

Chong, S. C. and A. Treisman (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, 45(7), 891–900.

Deroy, O. (2013). Object-sensitivity versus cognitive penetrability of perception. *Philosophical Studies*, 162(1), 87–107.

Hopkins, R. (1998). *Picture, Image and Experience.* Cambridge University Press: Cambridge.

King, A. (retrieved 06/09/2016) Aesthetic attitude, Internet Encyclopedia of Philosophy. http://www.iep.utm.edu/aesth-at/.

*Thanks to Casey O'Callaghan, Maarten Stenhaguen and Barry Smith for useful comments and discussions.

 © Mind Association 2017

Locher, P. J. (2012). Empirical investigation of an aesthetic experience with art. In Shimamura, A. & Palmer, S (eds) *Aesthetic Science: Connecting Minds, Brains, and Experience*, pp. 163–188. Oxford: OUP.

Locher, P., E. A. Krupinski, C. Mello-Thoms, and C. F. Nodine (2007). Visual interest in pictorial art during an aesthetic experience. *Spatial Vision*, 21(1), 55–77.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 84(1), 24–62.

Maynard, P. (2016). 'Neuroaesthetics', Gombrich, and Depiction'. *Brit J Aesthetics*, 56, 191–201.

Noe, A. (2011) Art and the limits of neuroscience. New York Times, Dec 4th, The Stone http://opinionator.blogs.nytimes.com/2011/12/04/art-and-the-limits-of-neuroscience/?_r=0.

Peacocke, C. (1987). Depiction. *Philosophical Review*, 96, 383–410.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(03)), 341–365.

Snowdon, P. (2010). On the what-is-likeness of experience. *The Southern Journal of Philosophy*, 48(1), 8–27.

Vogt, S. and S. Magnussen (2007). Expertise in pictorial perception: eye-movement patterns and visual memory in artists and laymen. *Perception*, 36(1), 91–100.

doi:10.1093/mind/fzw073 Advance Access publication 14 February 2017

Persons, Interests, and Justice, by Nils Holtug. New York: Oxford University Press, 2010. Pp. vi + 352.

Persons, Interests, and Justice (PIJ) develops and ties together two important areas of normative ethics—individual self-interest and distributive justice. Holtug makes important contributions in both areas, and writes with exceptional clarity and rigour. PIJ overlaps significantly with Derek Parfit's *Reasons and Persons* but goes even further in its investigation of the moral significance of individual self-interest, offering what is perhaps the most extensive and systematic exploration of the topic to date.

The book has three main parts. Part I (Chapters 1-4) focuses primarily on the topic of individual self-interest. It contains an extended discussion of what Holtug calls *The Identity View*, according to which one has a self-interest at time t_1 in some individual benefiting at some earlier or later time t_2 *if and only if* one is numerically identical to this individual; and the strength of the self-interest is strictly proportional to the size of the benefit (p. 26). Holtug rejects The Identity View (Chapter 3) by appealing to a hypothetical

 © Mind Association 2017

case of personal 'fission' in which person *a*'s cerebral hemispheres are divided and transplanted into separate bodies (p. 74). According to Holtug, when *a*'s cerebral hemispheres are divided and transplanted, *a* ceases to exist—he cannot be identified with either (or both) of his 'offshoots', *b* and *c*. Yet, Holtug claims, prior to his fission, *a* has a self-interest in *b* and *c* benefiting. For according to Holtug, what grounds one's future- or past-oriented self-interest is not numerical identity but *relation M*: the continuous physical realization of one's psychology (p. 99). Ultimately, Holtug endorses *The Prudential View*: one has a self-interest at time t_1 in some individual benefiting at some earlier or later time t_2 *if and only if* one, at t_1, is M-related to that individual at t_2; and the strength of the self-interest depends only on the size of the benefit and the strength of the M-relation (p. 104). Since people tend to change psychologically over time, typically one's self-interest in enjoying past or future benefits (units of welfare) will be *weaker* the further into the past or future these benefits lie. Relative to one's present perspective, such benefits are *discounted* due to the weakening of the M-relation between one's present self and one's past or future self.

Holtug considers how one's self-interest extends into the past or future when *more than one* past or future individual is M-related to one. He proposes *The Additive Prudential View*: one's self-interest is an additive function of the appropriately discounted benefits that befall what we might call one's *M-relatives*, i.e., those to whom one is M-related (p. 118). Holtug observes that since one could have many M-relatives, The Additive Prudential View leads to *The Repugnant Conclusion about Self-interest*: there is some finite number of individuals, all very weakly M-related to one, and all with lives that are equal in value to one's future, such that one has a stronger self-interest in these continuers coming about than in one's own survival (p. 122). Holtug thinks this conclusion is less counter-intuitive than it initially seems, since it may be 'difficult to fully fathom … how much value would be realized by an appropriately large number of continuers' (pp. 123-4).

Chapter 5 defends *The Value of Existence View*: one can be benefited (or harmed) by coming into existence (p. 130). In defence of this view, Holtug argues that if I have positive welfare then:

(1) my existence is better for me than my non-existence.

A worry about this claim is that it seems to imply:

(1) if I hadn't existed, this would have been worse for me.

Derek Parfit and John Broome have argued that (2) is incoherent, as it implies that *non-existing individuals* can stand in comparative relations. But according to Holtug, accepting (1) doesn't commit us to (2). Rather, he insists, the case for (1) can be made by comparing two states of affairs—*my existing* and *my not existing*—both of which *exist* at the actual world as

 © Mind Association 2017

abstract entities, although only the former *obtains*. Because *my existing* has positive value for me and *my not existing* has no positive value for me, Holtug argues, the former is better for me than the latter, and this fact establishes (1). Yet, it seems, bringing me into existence benefits me *comparatively* only if the alternative (my non-existence) would have been worse for me; and it is precisely this claim that Holtug must deny if his defence of (1) is to be successful.

Part II (Chapters 6-9) is about distributive justice narrowly construed as *axiology*—the part of normative ethics that concerns the evaluation of outcomes. For simplicity, Holtug sets aside his views about relation M and adopts a welfarist axiology according to which the value of an outcome is an increasing function *only* of the individual benefits it contains (p. 155). This axiology is consistent with total utilitarianism, egalitarianism, and prioritarianism, and much of Part II is devoted to comparing the respective merits of these three views. Chapter 7 presents a novel defence of *The Levelling Down Objection* (LDO) to egalitarianism. According to Larry Temkin, LDO assumes *a strong narrow person-affecting principle*: outcome *A* is *in no respect* better than outcome *B* if there is no individual for whom *A* is better than *B*. Temkin rejects this principle because it fails to adequately resolve *The Non-Identity Problem*. But Holtug argues that LDO can be established by appealing to what he calls *The Strong Wide Person-affecting Principle* (SWPP): outcome A is *in no respect* better than outcome B if both of the following are true: (a) if A obtained, there would be no individual for whom A was better than B, and (b) if B obtained, there would be no individual for whom B was worse than A (p. 186). If we level down, there will be no individual for whom the resulting outcome is better; and if we don't level down, there will be no individual for whom the resulting outcome is worse. SWPP therefore entails that the outcome of levelling down is in no respect better than that of not levelling down; and this suffices to establish LDO. Furthermore, SWPP avoids The Non-Identity Problem if one accepts Holtug's view that existence can be better for an individual than non-existence.

Chapter 8 defends *prioritarianism*: the value of an outcome is proportional to the sum of weighted individual benefits it contains, where benefits are weighted such that they gain a greater value the worse off the individual is to whom they accrue (p. 205). Unlike egalitarianism, prioritarianism focuses on absolute rather than relative welfare levels. Holtug prefers prioritarianism to egalitarianism because the former avoids The Levelling Down Objection and seems to better reflect our special concern for the worse-off.

Together with The Value of Existence View—the view that creating an individual can benefit (or harm) that individual—prioritarianism entails *The Repugnant Conclusion* as well as *The Super-repugnant Conclusion*: an outcome in which everyone's life is barely worth living can be better than another in which everyone's life is very good, even if the former has a lower total sum of welfare (p. 254). Holtug considers various ways of avoiding these conclusions, such as restricting the class of individuals whose welfare matters, and making

 © Mind Association 2017

outcome value a non-additive function of individual welfare. As Holtug demonstrates, these proposals face serious objections. Ultimately, he concludes, we should (i) accept the Repugnant and Super-repugnant Conclusions as corollaries of the most plausible (or least *im*plausible) axiology, or (ii) avoid these conclusions by restricting our axiology to comparisons of outcomes that contain the same number of individuals, or (iii) continue to search for a more plausible axiology that avoids the pitfalls associated with (i) and (ii). Holtug remains agnostic as to which of (i)-(iii) is the most plausible way to proceed.

In Part III (Chapter 10) Holtug emends prioritarianism in light of his Prudential View of self-interest developed in Part I. The result is *Prudential Prioritarianism*: the weaker one's self-interest at time *t* in the benefits that occur in one's life (or the lives of those to whom one is M-related), the greater the value of a further benefit to one at *t* or an increase in the strength of the M-relation that connects one at *t* to benefits that accrue to one (or those to whom one is M-related) at other times. Holtug's final view (his preferred axiology) is that the value of an outcome is an additive function of the individual contributions described in the above definition of Prudential Prioritarianism—where an individual's contribution is an additive function of the contributions she makes at particular times (p. 309).

As Holtug acknowledges, his final view is incompatible with the Strong Wide Person-affecting Principle (SWPP), since it implies that an outcome can be made better (or worse) through mere strengthening (or weakening) of the M-relation, which need not alter the distribution of individual benefits. Ultimately, Holtug replaces SWPP with *The Prudential Person-affecting Principle* (PPP): outcome A is *in no respect* better than outcome B if both of the following are true: (a) if A obtained, there would be no individual who, at some time, had a stronger self-interest in A than in B, and (b) if B obtained, there would be no individual who, at some time, had a weaker self-interest in B than in A (p. 336). As Holtug shows, PPP, like SWPP, supports a version of the Levelling Down Objection against egalitarianism.

Holtug's final view is fascinating, and it is the first proposed axiology to treat outcome value as a function of individual self-interest. However, it is not entirely clear how, according to this view, the fulfilment of individuals' self-interest determines the overall goodness of outcomes. For example, suppose that I undergo fission and that each of my 'offshoots' is fully M-related to me. Recall that according to *The Additive Prudential View*, which Holtug accepts, one's self-interest is an additive function of the appropriately discounted benefits that befall those to whom one is M-related. This implies not only that I have a self-interest in each of my offshoots benefiting, but that each of my offshoots has a self-interest in *my* benefiting. But then how, in this case, should the overall goodness of the outcome be determined? If outcome goodness is a function of self-interest fulfilment, should benefits that befall one of my fission offshoots be counted *twice*—once for their contribution to the fulfilment of the offshoot's self-interest and once for their contribution to the fulfilment of my

 © Mind Association 2017

self-interest? Should my benefits be counted *three times*—once for their contribution to the fulfilment of my self-interest and once for their contribution to the fulfilment of each of my offshoots' self-interest? If the answer is *yes* then the determination of outcome goodness seems to involve a kind of strange double-counting of benefits. But if the answer is *no* then, assuming that *The Additive Prudential View* is correct, outcome goodness does not fully reflect the fulfilment of individuals' self-interest.

Despite such lingering questions about the connection between self-interest fulfilment and outcome goodness, *Persons, Interests, and Justice* is an outstanding book filled with illuminating and insightful discussion of a wide range of topics.*

Institute for Futures Studies
timothy.campbell@iffs.se
TIM CAMPBELL

doi:10.1093/mind/fzw077 Advance Access publication 16 February 2017

*Funding from the Swedish Research Council is gratefully acknowledged.

Between Probability and Certainty: What Justifies Belief, by Martin Smith. Oxford: Oxford University Press, 2016. Pp. xi + 213.

It is fairly natural to suppose that having justification for believing that p is a matter of possessing evidence that makes p likely to be true. If this is right, then it would appear that one is justified in believing that one's lottery ticket has lost (after the drawing but before one hears of the announcement). And if this is also right, then one is also justified in believing, of each and every lottery ticket, that it has lost, from which one can deduce—and, assuming competent deduction preserves justification, presumably achieve justification for believing—that every ticket has lost! Since one is also very well justified in believing that one of the tickets has won, the idea that justification depends on evidence that makes one's belief highly probably true generates a paradox: S is justified in believing both that every ticket lost and that one ticket won.

Contrast the lottery problem with the Gettier problem. In a typical case, whether of the sort Gettier himself described or of the fake barns variety, there are elements of luck beyond the ken of the agent: good luck in achieving a true belief and bad luck due to features of the environment that undermine the agent's ability to know (such as the existence of fake barns in the vicinity of the real barn that is the object of the agent's perception and belief).

In Gettier cases but not in lottery cases, there exist factors unknown to the agent that bear on the agent's inability to know despite having justification. Thus in Gettier cases, there seems little scope for a negative assessment of the agent, whereas in lottery cases, following the evidence leads to inconsistent beliefs, something surely to be avoided. In addition, as intimated above, multi-premise closure (MPC) clearly plays a role in generating the

 © Mind Association 2017

inconsistent beliefs, but this creates a related problem for the risk-minimization conception of justification (RMC) briefly described in the first sentence of this review. Surely, it seems, one is justified in believing whatever one competently deduces from premises that one is justified in believing. Yet the intuitively plausible MPC strains the popular RMC. The small risk of error associated with each proposition of the form *that ticket i lost* snowballs as the number of premises of that form increases. In the lottery, it is quite clear that through risk aggregation, one's conclusion that no ticket won is not at all likely, given one's knowledge that there is a winner.

Perhaps this suggests that RMC is not a good model for justification, and maybe also that one is not even justified in believing, of one's own ticket, that it lost. Why, after all, do we think there is a marked difference in justification between one's believing that one's ticket lost because it is only one of a million tickets, and one's believing that one's ticket lost because the local newspaper said so, when we also know that the chances of a misprint in the paper are at least one in a million? If RMC were true, one would be less justified if one's belief were based solely on reading the results than if it were based on the fact that one holds only one ticket in a million. (This is a straightforward but not entirely correct way to compare the implications of RMC for these cases, for if there is a misprint, it does not follow that one's ticket has *won*. But the point is that RMC thinks of justification solely in terms, roughly, of the probability of the truth of the target belief given one's evidence, which generates problems like the one suggested in this paragraph.)

In his terrific and original book, *Between Probability and Certainty*, Martin Smith rehearses these problems for RMC. He argues that there is a fallibilist position that, as you may have guessed, lies between probability and certainty, one that maintains MPC. The basic position is that justification requires 'normic support' from one's evidence rather than simply support that increases the probability of the target proposition. Evidence normically supports a belief just in case 'it is necessary that all the most normal worlds at which [one's evidence] is true are worlds at which [the target belief] is true' (p. 42). A nice way to understand the relation of normic support is that, when one's belief that p is normically supported, there must be some explanation if one's evidence e is true but p is false (p. 40). Something *abnormal* has occurred. For example, if one's belief that one's ticket has lost is false but based on reading the paper, we need some explanation—perhaps there was some miscommunication or a typographical error—whereas if the false belief is based simply on the fact that one's ticket happened to win, no further explanation is necessary. That some ticket or other has won is not a departure from normal circumstances; that a typically reliable but fallible newspaper made a mistake arguably is such a departure.

I shall review some of the highlights of Smith's book with special emphasis (i) on further motivation for his view; (ii) on how his position nicely deals with problem cases and other examples; and (iii) on some of the welcome

 © Mind Association 2017

implications of Smith's formal framework. In my view, even some of the somewhat counterintuitive upshots of Smith's position, such as that one is *not* justified in believing that one's ticket lost, are defensible, perhaps even plausible.

Smith's view that a belief is justified just in case it is normically supported by evidence is motivated substantially by problems with RMC, principally its inconsistency with MPC and its inability to distinguish what one might call probative evidence from mere statistical evidence. I briefly considered the lottery case above, where RMC predicts that one is justified in believing that one's ticket has lost, and, if MPC is valid, that each and every ticket has lost, whereas the theory of normic support implies that one is not even justified in believing that one's ticket has lost. The related preface paradox also illustrates RMC's inconsistency with MPC and its failure to distinguish merely statistical evidence from (again, perhaps this is controversial, but so be it) genuinely probative evidence. One is justified in believing, of each claim made in one's own book, that it is true. By MPC one is justified in believing that every claim is true, whereas one announces, with apparent justification, that there is bound to be a false assertion somewhere in the book. Here, however, the theory of normic support generates a result different from that in the lottery case: one is (or at least could be) justified in believing that each claim is true, hence by MPC that every claim is true. This is because, for each claim made, given the author's evidence (and assuming that the author is meticulous), if it is false, there must be some explanation of that fact. Thus one is not justified in believing that at least one claim is false. In the lottery case, one's belief that one's ticket has lost is highly probably true given the evidence (/odds), but not justified. In the preface case, one's belief that all one's claims are true is highly unlikely but justified. Some will balk at this implication. It is perhaps the most controversial upshot in Smith's book. But it is certainly not a crazy position, and to mitigate its counterintuitive feel, I note that Smith's position is compatible with the idea that one is justified in believing that it is *probable* that there is at least one mistake in the book (for this is true in most normal worlds!). (See, for example, the brief discussion on pp. 83-4.) Below we shall consider other cases that lend support to Smith's normic theory of justification.

A selling point in favour of Smith's normic support theory is the fact that it is consistent with MPC, since 'if a body of evidence normically supports each of a set of propositions, then it will also normically support any proposition that they jointly entail' (p. 44), without generating the lottery and preface paradox because 'no body of evidence could provide simultaneous normic support for each of an inconsistent set of propositions' (p. 45). But why not reject MPC, as is becoming increasingly fashionable, rather than accept normic support? In Chapter 4, Smith demonstrates that rejecting MPC isn't merely counterintuitive, but comes with other costs—principally, that one must also deny either *cumulative transitivity*, *strengthening*, *cautious*

 © Mind Association 2017

monotonicity, or single premise closure. (The following are Smith's definitions. *Cumulative transitivity*: 'If E provides justification for believing P and E ∧ P provides justification for believing Q then E provides justification for believing Q' (p. 81). *Strengthening*: 'If E ∧ P provides justification for believing Q, then E ∧ P provides justification for believing P ∧ Q' (ibid.). *Cautious monotonicity*: 'If E provides justification for believing P and E provides justification for believing Q then E ∧ Q provides justification for believing P' (p. 82).) Of course, anyone who rejects MPC may well be willing also to reject one or more of these intuitively valid principles, but it is to Smith's credit that he shows that one would *have* to. In Chapter 7, Smith deftly walks us through elegant proofs that cumulative transitivity and cautious monotonicity are valid inference patterns for normic conditionals. This comes after Smith offers, in Chapter 6, a 'normal worlds' analysis of normic conditionals (for example 'If p were the case then normally q would be the case') that capture normic support relations, *à la* possible worlds analyses of counterfactual conditionals.

Another line of support and motivation for Smith's notion of normic support comes from his assumption that the goals of achieving knowledge and of having justification are normatively coincident—that 'there is no way to aim for one without automatically aiming for the other' (p. 8). According to RMC, one might well aim for (and achieve) justification for believing that one's lottery ticket will lose, even though one cannot aim for knowledge that one's ticket will lose. Smith seems to suggest that violating the Normative Coincidence Constraint (NCC) partly diagnoses the problems with RMC. Unfortunately, Smith isn't himself prepared to give a full-blown defence of NCC, and because it is fairly controversial, perhaps beginning the book with a chapter devoted to NCC isn't the most strategic approach. Unfortunately, too, the chapter is internally problematic. It begins 'with an unqualified assumption' (p. 8), but Smith never tells us what it is. I had to read the first page of Chapter 1 several times before deciding that he must be talking about NCC. (This is a very specific example where the reader might be confused, but more generally, while the book hangs together quite well in terms of thematic content, it is not always clear why the chapters and topics are arranged the way they are. The book would be more reader-friendly if there were greater use of sign-posting and perhaps an introduction that offers at least a $.50 tour of its contents.) At the end of the chapter, Smith seems to admit that he hasn't yet given much of an argument against RMC by appeal to NCC, but goes on to say that his aim is merely to give 'a glimpse of how some parts of epistemology might end up looking if we took [NCC] seriously' (p. 27). It may have served Smith's overall purposes to take up the question of NCC later in the book, after having given his best, considered arguments for normic support and against RMC. Alternatively, Smith might have simply offered NCC at the beginning of the first chapter not as an unqualified assumption but as a mere conjecture to investigate.

 © Mind Association 2017

Smith also demonstrates how his normic theory handles other cases in the literature. Hawthorne (2003) generates another kind of lottery paradox, couched not in terms of justification but of knowledge. I may know that I cannot afford a safari and know that this entails that I won't win the lottery, but I do not know that I won't win the lottery. For Smith, the analogous problem is that I am justified in believing I cannot afford a safari, and I know that this entails that I won't win the lottery, but as we've seen, I'm not justified in believing that I won't win the lottery. The normic theory, however, implies that I am not justified in believing I cannot afford a safari, because the world where I win the lottery is a normal one and in that world I can afford a safari.

Now, however, there is a new threat: scepticism. If I am not justified in believing that I cannot afford a safari because it implies another proposition that I am not justified in believing, viz, that I won't win the lottery, then am I not justified, for example, in believing that my team lost because that they lost implies that the reported result that I saw on ESPN's site is correct, which is something that I haven't independently verified? After all, the chances that the result is incorrectly reported are presumably better than my chances of winning a massive lottery, so if I am not justified in believing that my lottery ticket lost, it would seem that I am not justified in believing there is no error in the reported game result, and thus not justified in believing the report itself. Smith's theory denies this superficial parity. I am justified in believing that my team lost, because, again, if the result were wrongly reported, there would have to be some explanation—a misprint, or an inattentive reporter; it is not normal, unlike the normal (but unlikely) world where I win the lottery.

Chapter 5 aims for an account of comparative justification—of how one belief can be more justified than another. A potentially unwelcome implication of Smith's baseline normic theory of justification is that if p enjoys greater normic support than q, then p is categorically justified. Smith then sketches a version of his theory that would block this result, according to which categorical justification requires meeting a minimum threshold of degrees of normic support. The downside to this approach, one that afflicts RMC also, is that any particular threshold will seem arbitrary. One option here, suggests Smith, is to think of the relevant threshold as determined by practical interests, rather like the subject-sensitive invariantist theories of Hawthorne (2003) and Stanley (2005). Following this approach also permits a response to a particular concern one might have with Smith's diagnosis of the preface paradox. There seems to be a close connection between justified belief and grounds for action. What if the publisher were to charge the author $100,000 if the author asserts that every claim in the book is true and subsequently a false claim in the book were to come to light? Is the author still justified in believing that every claim in the book is true? If the threshold for justification depends on practical interests, then, given the high stakes, the

 © Mind Association 2017

author may not meet that threshold and in that case should refrain from asserting that every claim in the book is true.

Allow the preceding to serve as an overview of Smith's normic theory of justification. The book offers much more food for thought in an exploratory spirit, setting an agenda for fruitful research in many areas of interest to contemporary epistemology and its applications. *Philosophy of Law.* The theory of normic support can explain why specific testimony is a legitimate basis for legal findings whereas general statistical evidence is not, even when a finding is more probable given the statistical evidence than it is given the testimony (pp. 36ff.). *Contextualism.* Smith offers a contextualist account of justification modelled on contextualism about the truth conditions of knowledge attributions (§6.3). *Safety.* Smith shows how to combine a safety account of knowledge ('protection from error that could have easily obtained' (p. 132)) that coheres nicely with a normic theory of justification ('protection from error that could have normally obtained' (ibid.)). *Sensitivity.* One diagnosis of why we do not know a lottery proposition is that it is not believed sensitively—if it were false, we would believe it anyway (bracketing evidence of the lottery result). But sensitivity cannot explain how we can have justified *false* beliefs. (To be fair, sensitivity theorists don't intend to illuminate justification, but the point holds.) Suppose that the announcement of a lottery result is false. One would still be justified in believing the announcement, since in most normal worlds the announcement is true (pp. 6of.). *Knowledge Account of Assertion.* Smith's theory provides a plausible alternative to the knowledge account's explanation for why it is inappropriate to assert that one's ticket is a loser, according to which the reason is that one does not know, and one should assert only what one knows. For Smith, one is not even justified in believing that one's ticket is a loser. *Internalism and Externalism.* Smith presents a version of normic support that captures key insights of internalism about justification while maintaining a conception of the link between belief and truth that is essential to reliabilism (§6.4) (though perhaps reliabilists won't be satisfied with the nature of the belief-truth link that Smith articulates). *Degrees of belief.* Smith proposes a way of modelling degrees of belief based on a ranking-theoretic framework. (See, for example Huber, 2009.) The model presents an alternative to the more familiar Bayesian conception of credences.

I end the discussion of specific features of Smith's book on the issue of degrees of belief because it is here that Bayesians, elaborating RMC, can claim their greatest competitive advantage over alternative theoretical structures. Smith is careful not to denigrate or devalue the insights of a probabilistic approach to justification; indeed, he frequently characterizes his normic theory simply as an alternative, nowhere, so far as I could find, claiming to have established that it, and not RMC, is the *truth* about justification. And yet, generalizing the insights from the lottery problem, Smith argues that to believe solely on probabilistic terms 'can seem a clumsy, ill-fitting attitude, no

 © Mind Association 2017

matter how close to certain P might be' (p. 195). There seems to be something arbitrary about singling out 'the negation of P to dismiss' when 'other possibilities are, at best, on an epistemic par with it' (ibid.). On the other hand, 'If P is normically supported … belief no longer seems such a peculiar attitude. Singling out the negation of P to dismiss, far from being arbitrary, now seems to be the attitude that most faithfully reflects my epistemic situation' (ibid.).

These claims are pregnant. While some may be disappointed that Smith does not offer much detail about what constitutes the normal worlds that provide the basic framework of normic support, he is clearly on to something in taking this approach. In my view (perhaps more germanely, in my own experience reading the book), it does not take a lot of rethinking one's own assumptions to see that there's something right (even if one doesn't *fully* agree) in saying that one is not justified in believing a lottery proposition, and even that one *is* justified in believing, for example, that all the claims one makes in a book are true. One does not have any evidence that *this* ticket, as opposed to that one, will lose. One may have convincing evidence for every claim one makes in a book. I think, too, that even though we are not given a developed theory of normal worlds, Smith's claim that, when a normically supported belief is false, we expect there to be some explanation of how it came to be false, is an illuminating first step in crafting such a theory. Furthermore, the formal implications of normic support that Smith proves, dovetail neatly with what one would expect from the concept of justification.

One mark of good philosophy, it seems to me, is that not everyone will agree with its central tenets, and yet even dissenters will be able to see why somebody would hold them. Much of analytic philosophy is still, even since the naturalistic turn, conceptual business, and getting people to change their minds sometimes means trying to get them to rethink their basic conceptions. Philosophers often balk because they sense that they're being asked to go in for extensive revisionism. Smith asks us to rethink some commonly held assumptions about the concept *justification*, but interestingly, in this case I would argue that his basic notion of normic support takes us closer to a pre-theoretic concept of justification rather than further away, if the starting point is in the RMC family. (I am tempted to draw the broader moral that RMC represents a version of the naturalist ploy to reduce normative notions to merely probabilistic and statistical (or at any rate non-normative) ones, with predictable trouble at just the juncture where the notions were to meet, but that is a topic too big for this review, so perhaps this parenthetical gesture serves my purpose well enough.) Indeed, normic support seems tailor-made to the project of understanding justification.

Then again, that it is tailor-made lends itself to the concern that normic theory is, in some sense, uninformative. Perhaps it constitutes merely one way of unpacking the concept *justification* which, if correct, tells us little more than we already knew if we already had an adequate grasp of the concept. If

 © Mind Association 2017

that is your concern, fear not! Smith's book provides a highly original account of perhaps the central concept in epistemology, one that offers fertile ground for exploring just about every conceivable topic and idea of central interest to the field, new and old. I suspect that those who are newly introduced to the theory of knowledge will take to it like moths to a flame, and I hope that veterans will set aside their preconceptions and take it seriously as a force to be reckoned with.*

University of New Mexico KELLY BECKER
kbecker@unm.edu

References

Hawthorne, J. 2003. *Knowledge and Lotteries* (Oxford: Oxford University Press).

Huber, F. 2009. 'Belief and Degrees of Belief', in F. Huber and C. Schmidt-Petri eds. *Degrees of Belief* (Dordrecht: Springer).

Stanley, J. 2005. *Knowledge and Practical Interests* (Oxford: Oxford University Press).

doi:10.1093/mind/fzx002 Advance Access publication 14 March 2017

* My thanks to Martin Smith for suggesting clarifications on a couple of points in this review in response to a previous draft.

 © Mind Association 2017

GTU Library
3 2400 00823 8515

DATE DUE