

Being in Time

Advances in Consciousness Research (AiCR)

Provides a forum for scholars from different scientific disciplines and fields of knowledge who study consciousness in its multifaceted aspects. Thus the Series includes (but is not limited to) the various areas of cognitive science, including cognitive psychology, brain science, philosophy and linguistics. The orientation of the series is toward developing new interdisciplinary and integrative approaches for the investigation, description and theory of consciousness, as well as the practical consequences of this research for the individual in society.

From 1999 the Series consists of two subseries that cover the most important types of contributions to consciousness studies:

Series A: Theory and Method. Contributions to the development of theory and method in the study of consciousness; Series B: Research in Progress. Experimental, descriptive and clinical research in consciousness.

This book is a contribution to Series B.

For an overview of all books published in this series, please see

<http://benjamins.com/catalog/aicr>

Editor

Maxim I. Stamenov

Bulgarian Academy of Sciences

Editorial Board

David J. Chalmers

Australian National University

Axel Cleeremans

Université Libre de Bruxelles

Gordon G. Globus

University of California Irvine

Christof Koch

California Institute of Technology

Stephen M. Kosslyn

Harvard University

Steven Laureys

University of Liège

George Mandler

University of California at San Diego

John R. Searle

University of California at Berkeley

Petra Stoerig

Universität Düsseldorf

Volume 88

Being in Time. Dynamical models of phenomenal experience

Edited by Shimon Edelman, Tomer Fekete and Neta Zach

Being in Time

Dynamical models of phenomenal experience

Edited by

Shimon Edelman

Cornell University

Tomer Fekete

Stony Brook University

Neta Zach

University of Pennsylvania

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Being in time : dynamical models of phenomenal experience / edited by
Shimon Edelman, Tomer Fekete, Neta Zach.

p. cm. (Advances in Consciousness Research, ISSN 1381-589X ; v. 88)

Includes bibliographical references and index.

1. Phenomenological psychology. 2. Experience--Psychological aspects. 3.

Consciousness. 4. Time. I. Edelman, Shimon. II. Fekete, Tomer. III. Zach, Neta.

BF204.5.B45 2012

153--dc23

2012016358

ISBN 978 90 272 1354 9 (Hb ; alk. paper)

ISBN 978 90 272 7359 8 (Eb)

© 2012 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Introduction	VII
Time after time: Temporality in the dynamic brain <i>Dan Lloyd</i>	1
Neuronal reflections and subjective awareness <i>Rafael Malach</i>	21
From probabilities to percepts: A subcortical “global best estimate buffer” as locus of phenomenal experience <i>Bjorn Merker</i>	37
Being in time <i>Shimon Edelman & Tomer Fekete</i>	81
The (lack of) mental life of some machines <i>Tomer Fekete & Shimon Edelman</i>	95
Restless minds, wandering brains <i>Cees van Leeuwen & Dirk J.A. Smit</i>	121
Fuzzy consciousness <i>Stephanie Huettenlocher & Michael Spivey</i>	149
Two dynamical themes in Husserl <i>Jeff Yoshimi</i>	165
Desiderata for a mereotopological theory of consciousness: First steps towards a formal model for the unity of consciousness <i>Wanja Wiese & Thomas Metzinger</i>	185
The brain and its states <i>Richard Brown</i>	211
An integrative pluralistic approach to phenomenal consciousness <i>Rick Dale, Deborah P. Tollefsen & Christopher T. Kello</i>	231
Index	259

Introduction

“The relation of experience to time has not been profoundly studied. Its objects are given as being of the present, but the part of time referred to by the datum is a very different thing from the conterminous of the past and future which philosophy denotes by the name Present. The present to which the datum refers is really a part of the past – a recent past – delusively given as being a time that intervenes between the past and the future. Let it be named the *specious* present, and let the past, that is given as being the past, be known as the obvious past. All the notes of a bar of a song seem to the listener to be contained in the present. All the changes of place of a meteor seem to the beholder to be contained in the present. At the instant of the termination of such series, no part of the time measured by them seems to be a past.”

(William James, *The Principles of Psychology*)

Intuition has it that one’s phenomenally experienced existence is confined to a point in time – the present moment, as it were, at least on those rare occasions when the mind manages to refrain from wandering in time and space to virtual worlds of its own making. Although the conception of an instantaneous present to which experience is confined is readily dispelled by the kind of analysis that prompted William James to call it “specious,” it is still acceptable for a theory of consciousness to be summarily exempted from addressing the issue of the temporality of phenomenal experience. The chapters comprising this volume represent a collective attempt on the part of their authors to redress this metatheoretical anomaly, by exploring the implications of the idea that phenomenal experience is best understood – indeed, can only be understood – as being in time.

The exclusive theoretical focus on temporally punctate events, such as the vaunted instantaneous phenomenal percept, which is induced by the onset of an external stimulus and which “enters consciousness,” after some delay, at a definite moment in time, is typically accompanied by a tacit assumption that our stream of consciousness (curiously enough, another Jamesian coinage) plays out like a movie. Movies, of course, consist of series of static snapshots that follow each other at short enough intervals so as to appear continuous. The analogy between conscious experience and a movie is, however, problematic, in that it puts off the temporal continuity issue rather than resolving it. Indeed, the seamlessness of motion picture perception does not inhere in the collection of static images that comprise it, but rather is the outcome of a laborious process that is

not felt by the viewer as such only because his or her brain happens to be good at the kind of spatiotemporal interpolation that underlies the illusion of motion continuity.

Given that letting our experience be up to someone else would make no sense, a theory of experience must be *intrinsic* to the experiencer. In particular, it may not posit anything like the interpretation of the experiencer's brain activity by an external observer. Thus, a "cinematographic" theory of experience must include an explanation of how the brain juggles the dual tasks of (i) generating the series of frozen states corresponding to the phenomenal moments and (ii) processing the succession of its own states so as to imbue its experience with temporality. A radical alternative to this dual-task approach would be to identify experience with time-extended, continuous, dynamical brain processes to begin with, rather than assuming that continuity stems from self-observation. The four chapters in the first part of this book outline these competing theoretical approaches and examine their computational and neurobiological implications.

Part I. The dynamics of experience: Possible brain mechanisms and computational considerations (Lloyd; Malach; Merker; Edelman & Fekete)

"Time is the substance of which I am made. Time is a river that sweeps me along, but I am the river; it is a tiger that mangles me, but I am the tiger; it is a fire that consumes me, but I am the fire."

(Jorge Luis Borges, *The New Refutation of Time*)

Our conscious experience results to a large extent from our brain's interaction with the environment. Both brain activity and the consciousness that it gives rise to change in time. Therefore, the relationship between changes in brain activity and changes in phenomenal experience are crucial for understanding and modeling the latter.

Lloyd, who advocates a dynamical systems approach to these issues, introduces it by the way of Husserl's phenomenology, which stresses both the fundamentally flowing aspects of experience and its invariant aspects, such as the persistence of objects and entities. The dynamical systems perspective emerges as a middle way between those phenomenological frameworks that neglect the stationary on the one hand, and on the other hand the traditional cognitive sciences that neglect the temporal.

Under this perspective, an experience is a region in neural activation space in which the trajectories associated with a type of experience reside. As an example, Lloyd looks for the signature of such phenomenal structure in fMRI data from a virtual reality experiment that consisted of interlaced driving, passive

viewing, and fixating epochs. Using independent component analysis (ICA) and multidimensional scaling (MDS), he shows that brain activation states cluster by epoch type. As revealed by hierarchical cluster analysis, brain activation is structured at multiple scales: as the number of clusters increases, progressively finer temporal structure in activity manifests itself. At such finer scales, the analysis resolves the brain's coherent wandering through activation space, which is seen to slow down during fixations.

Multiscale analysis thus brings out both the stable brain states sought after by cognitive sciences and the dynamic ebb and flow of brain activation expected from phenomenology. Inspired by the tri-partite phenomenology of Husserl, according to which the present enfolds in it elements of the past and the future, Lloyd uses machine learning techniques to show that each state can indeed be used to predict following and preceding states. As one would expect, the accuracy of such prediction decreases, the further away one is from the reference “now” moment. A simulated “lesion analysis” suggests that this temporal structure is spatially distributed in the brain – again, as expected from a putative substrate of a fundamental facet of experience. Lloyd ends his chapter by noting that in a successful theory of consciousness higher order properties must be explicitly modeled and suggests that explicit representations of activation-space trajectories may underlie many such properties.

Malach's chapter proposes an account of phenomenal awareness in terms of the processing of ambiguous sensory inputs. Malach argues that the ambiguity is resolved (sometimes improperly, as in the case of visual illusions) through interaction between the input and information already stored in the cortical visual system. The process of ambiguity resolution is hierarchical, with most anterior regions of the cortex giving rise to the full holistic percept. The mechanism whereby the input interacts with the hierarchy of “templates” is identified with recurrent activation of small neural assemblies, or local neuronal “ignitions.”

In support of this explanation, Malach cites experimental evidence of high firing rates, sustained activity, and dense local connectivity – three prerequisites of the ignition account that are indeed found in cortical circuits. A key prerequisite – evidence of a tie-in between conscious awareness and local reverberatory activity – is, however, still heavily debated. The local ignition hypothesis will, therefore, stand or fall, depending on whether or not such evidence solidifies.

The account of phenomenal awareness developed by **Merker** in his chapter also hinges on the observation that sensory information is inherently ambiguous and that, correspondingly, its processing by the brain must be probabilistic and must eventually resolve the uncertainty, given that typical perceptual states are unambiguous. This resolution process is posited to occur outside consciousness, with only its outcome becoming phenomenally available, in the

form of the unambiguous global best estimate of the current scene as a whole. On Merker's account, the uncertainty resolution is orchestrated by subcortical circuits, which complement in this respect the inherently probabilistic *modus operandi* of the cortex.

The disambiguation process is complicated by the time constraints to which it is subject. The funneling of massive amounts of information into a unified flow of consciousness requires constant buffering. Only when this information is integrated into a holistic, dynamic, ego-centric model of body-world interactions does the mosaic of representations become conscious. The buffer thus serves to integrate the disambiguated stream of probabilistic data, creating the specious present. Merker shows that fitting candidates for this function are the higher order nuclei of the dorsal thalamus, specifically, the caudal reaches of the dorsal pulvinar.

Edelman and Fekete discuss some of the constraints that a computational model of experience must satisfy if it is to deal properly with time. Specifically, they focus on the interaction between the computational requirements of tractability and timeliness on the one hand and, on the other hand, the fundamental characteristic of phenomenality: its autonomy, that is, the requirement that putative theoretical constructs in their entirety be intrinsically computable by the brain itself, without recourse to an external observer. Edelman and Fekete argue that the common move in the sciences of the mind – equating mental content with the *instantaneous* state of the brain – is fundamentally flawed: due to physical constraints such as speed of communication and the computational intractability of asynchronous distributed consensus, an instantaneous state can only be a figment of the external observer's imagination.

In comparison, the *dynamics* of the system – that is, its trajectory through the space of its possible states – is no more and no less than what the system actually computes. Consequently, trajectories through a system's state space seem to be the only viable theoretical primitive for modeling experience (cf. Wiese & Metzinger this volume). The theoretical benefits of this stance are immediate, and include accounting for the role of silent units in establishing concurrent experience, the ability to intrinsically (and counterfactually) define the confines of a single system, and of course the possibility of offering an intrinsic account of the time-infused essence of experience. In this light, it becomes clear that the popular notion of attractor dynamics as a building block of cognition (and of consciousness) is flawed: outside of strict stimulus-response scenarios within the confines of a laboratory it is simply not applicable. In contrast, a richer conception of dynamics, such as chaotic itinerancy, which uses properties of trajectory spaces as explanatory primitives, seems to be up to the task of explaining being in time.

Part II. The dynamics of experience: Theoretical implications (Fekete & Edelman; van Leeuwen & Smit; Huette & Spivey)

“Every instant is autonomous. Not vengeance nor pardon nor jails nor even oblivion can modify the invulnerable past. No less vain to my mind are hope and fear, for they always refer to future events, that is, to events which will not happen to us, who are the diminutive present. They tell me that the present, the “specious present” of the psychologists, lasts between several seconds and the smallest fraction of a second, which is also how long the history of the universe lasts. Or better, there is no such thing as “the life of a man,” nor even “one night in his life.” Each moment we live exists, not the imaginary sum of those moments.”

(Jorge Luis Borges, *A New Refutation of Time*)

Is every instant indeed autonomous, as Borges would have it? Is there not a bridging process that links together the instances into a chain that is the flow of consciousness? This part of the book looks at how the flow of time in phenomenal experience is addressed by several dynamical computational models. In these models, phenomenal awareness is modeled in terms of the system’s state-space trajectories, with a special focus on the structure of this space.

Fekete and Edelman examine the ability of one class of such spaces, those that arise in digital computer systems, to give rise to phenomenal experience. They begin by analyzing the familiar digital replacement scenario – a thought experiment in which the brain is replaced, neuron by neuron, with digital chips. A careful, physiologically informed analysis of this scenario reveals that the purported argument for the equivalence of the resulting digital “brain” to the original – the principle of organizational invariance (OI) – falls short on two counts. First, it relies on an inadequate mode of abstraction, which only insists on preserving input/output relations when replacing a functional unit such as a neuron. Second, it depends critically on the notion of sufficient grain for modeling (in space and time) – an arbitrary move that is shown to result in inconsistency.

Fekete and Edelman next analyze the possibility of computer-simulated neural networks as a putatively sufficient substrate for experience. They show that if implementation details are properly considered, the claim that a digital computer realizes a particular dynamical system as specified by a given set of equations is essentially a matter of external attribution, thus failing the autonomy requirement of theories of experience. Among the reasons for this failure are the inability of finite discrete systems (as opposed to analog systems) to intrinsically represent numbers, and the lack of physical (and hence ontological and epistemological) distinction within a simulation between dynamical variables and causal interactions/forces that operate on those variables, as both are represented (instantiated) in the exact same format (bits). Further analysis suggests that the minimal requirement for a machine to be capable of realizing experience is to be an open, analog, time

continuous dynamical system – unlike digital computers, which fail to capture the intrinsic temporality of consciousness.

Van Leeuwen and Smit's chapter puts forward the hypothesis that the mental processes occurring in the common situation where the perception is underdetermined by sensory data are strongly affected by mind wandering, which makes perception more dependent on past history and variable over time. This view, which may be compared to that of Lloyd (this volume), stands in contrast to models of perceptual processing that focus exclusively on invariance. Much like Merker and Malach (this volume), they see the importance for this process in giving the perceptual system flexibility. The posited dynamical process underlies consciousness but is not conscious in itself and therefore should be examined by psychophysics, not introspection.

To study this process, Van Leeuwen and Smit sought patterns in the dynamics of perceptions reported by subjects who viewed Necker cube stimuli, which they augmented with EEG data. In addition, they also looked at the subjects during rest, and they were able to identify patterns of synchrony between different parts of the brain, in the alpha frequency band of the EEG potential. When the observers were engaged in a task, these patterns were replaced by synchrony events in the beta and gamma bands. Van Leeuwen and Smit suggest that the alpha-band activity corresponds to non-conscious wandering that leads up dynamically to the emergence of experienced, phenomenal states.

Huette and Spivey argue that before consciousness can be understood, it must be modeled mathematically. In our everyday life we experience a stream of consciousness that is always an admixture of concepts, percepts, and emotions, all of which are highly interactive, content-sensitive and fleeting. That none of these are experienced purely in isolation suggests that consciousness should be modeled as a fuzzy dynamical system, in which experience is equated with the trajectory though a mental state space. This space is populated by entities such as concepts and percepts – regions with attracting properties. When experience – the system's trajectory – passes near an attractor, the corresponding concept becomes part of it, taking the form of a function of thousands of variables that span multiple time scales.

Huette and Spivey draw a distinction between consciousness, or being in time, and awareness, which consists of concepts and percepts. In the process of maturation, the brain learns to constrain the impact of the environment so as to form an increasingly clustered mental space (clusters being attracting regions), which results in increasingly structured trajectories – that is, phenomenal flow. In this fuzzy representation space, the most travelled paths between concepts form “cylindrical” sets – bundles of trajectories that over time acquire attracting properties of their own. The distance of the trajectory to the core of a concept is what

determines awareness. In early stages of development, these cylinder sets, representing structured thought or action, are sharply delimited and distinct, and so is awareness. With time, the trajectories tend to “straighten out” as shortcuts result from striving for efficiency. Thus, the authors predict that consciousness, which is best understood as being in time, strengthens with life experience, as mental space becomes more clustered and cylindrical sets more well formed, while awareness peaks at a young age, then declines, as mental economy takes over.

Part III. The dynamics of experience: A philosophical perspective (Yoshimi; Wiese & Metzinger; Brown; Dale, Tollefsen & Kello)

“No man ever steps in the same river twice, for it’s not the same river and he’s not the same man.”

(Plato, *Cratylus*, ascribed to Heraclitus)

The last part of the book offers several philosophical and methodological perspectives on the core questions that have been brought up so far. Does a model have to have a temporal aspect in order to accurately capture consciousness? Or perhaps a set of stimulus-response associations can describe consciousness reasonably well, at least in some cases?

Yoshimi sets out to illustrate how Husserl’s theory of consciousness is naturally expressed in terms of dynamical system theory, and hence offers a rich source of ideas for the empirical study of experience. He suggests that the basic construct, from which all other Husserlian constructs can be derived, is that of a phenomenological state space – the space of all possible conscious states C . The stream of consciousness is then a succession of states in C . The question arises as to what might be the principles according to which C is organized. According to the mereology – calculus of parts and wholes – proposed by Husserl, C can be broken down into subspaces, corresponding, e.g. to aspects of experience such as vision and sound (cf. Wiese & Metzinger this volume). Our experience of external reality then results from an inner model (horizon) of the world, in which our knowledge, our past (immediate and distant), and our dispositions coalesce to constrain (“constitute”) possible phenomenal states.

This model, the horizon, is in fact a structure in C , which varies in its degrees of articulation: regions reflecting paucity of experience are less articulated, but can become more so with learning. These observations can be formalized in terms of a three-argument “expectation function,” which maps a person’s overall background knowledge, bodily movements, and current visual experience (say) into a probability distribution on V (a subset of C). This function determines the degree to which unfolding events feel surprising or familiar to us, and hence

leads to elaboration of the horizon to the extent we are surprised. From a geometric standpoint, it can be thought of as “highlighting” a structure in *C* holding the possible experiences, given a context. Husserl tried to offer a more detailed analysis along these lines in terms of transcendental-*eidetic* “laws” – invariances that enable experience in various domains (e.g. possible color experiences, conversations, etc.). Such domains are structures in *C*, and can be probed using “free variation” to discover the rules that govern them, by varying an exemplar pertaining to the domain, so as to infer invariance. Yoshimi discusses several examples of this principle, such as the *eidetic* laws regarding the experience of movement, and describes his research in applying these notions to the study of embodied neural networks.

Wiese and Metzinger catalogue the analytical building blocks for a phenomenological theory of conscious experience. They explore the complex relationships between unity and temporal continuity found in experience, by the way of multiple constraint satisfaction across different levels and modes of analysis. The constraints derived from unity and temporal continuity involve mereotopological notions (e.g. parthood from mereology and openness from topology). The authors suggest that mereotopology is a fruitful theoretical approach that can inform empirical investigation, in particular enabling integration of data and theory at different levels. As a case study, they describe a critical analysis of the notion of “duration of experience.” Finally, the authors propose a representational architecture that affords a theoretical framework for understanding the unity of consciousness. In this architecture, a representational system comprises subsystems representing objects, but also the representational activity of the other subsystems (and through that, their own activity). This dynamic coupling of local and meta representations leads to representational fusion, which underlies the global unity of experience. Wiese and Metzinger argue that this globally coherent temporally extended state corresponds to the phenomenon of the specious present.

Brown begins by noting that several types of consciousness have been identified by research: creature consciousness (a creature being awake and responding to stimuli), transitive consciousness (being conscious of things and events), state consciousness, access consciousness, and phenomenal consciousness. He suggests that creature consciousness can be understood in terms of the AIM model (Hobson et al. 2000), according to which the global chemical state of the brain (levels of neuromodulation) realizes a low dimensional state space in which deep sleep, dreaming, and wakefulness occupy distinct regions (as do mood disorders, e.g. depression). On this view, creature consciousness is a state of the brain, while a conscious state, being a particular, is a brain state.

Brown then goes on to argue that all other forms of consciousness (brain states) can be reduced to transitive consciousness, which is understood in terms of synchronized neural activity, with high frequency synchrony being essential to the brain's representational vocabulary (cf. Malach this volume), and low frequency synchrony serving as a carrier to disseminate information between brain regions (cf. van Leeuwen & Smit this volume). He claims that experimental evidence supports the Higher-Order Representation Of a Representation (HOROR) theory of phenomenal consciousness, which shares similarities with both lower and higher order theories. As evidence for this idea, Brown cites the impact of the acquisition of new concepts on experience (e.g. in wine tasting), suggesting that all phenomenal content derives from concepts and categories. Another line of supporting evidence comes from experiments indicating that in detection tasks subjects report unattended stimuli as more visible than attended ones with the same first-order representational strength (as quantified by signal detection sensitivity). This leads to a claim that the higher order representation in itself suffices for experience, resulting in a kind of physical identity theory. Brown dismisses some of the familiar arguments against this move based on conceivability (e.g. zombie arguments), by applying them to physical properties, which turn out to be equally susceptible to such an attack. Thus, as it is synchrony that individuates states of consciousness (states of the brain), and as synchrony is a temporally extended property, experience is seen as inherently being in time.

In the closing chapter, **Dale, Tollefsen, and Kello** review and try to integrate current consciousness research and theory. They identify several distinct avenues through which the problem has been approached and argue that as consciousness is a complex phenomenon, it will require a pluralistic explanation. In particular, the multiscale property of phenomenal experience allows numerous theories to have each an explanatory role, depending on one's meaning of "experience," and on the measurement grain size chosen for the analysis. The authors argue that such a multiscale account of experience is best articulated through the language of dynamics and complexity science. They suggest that explanatory coherence will be achieved through investigation of the temporally extended facets of experience expressed in its sustained nature, which is nevertheless constituted by interacting parts (content) at multiple scales in space and time. Moreover, they argue that the patterns of complexity found at one level – such as that of neuronal dynamics – cannot explain consciousness, as it recurs in various unrelated physical phenomena. Rather, it is the unique multiscale organization of the dynamics introduced by coupling with our physical and social environs, resulting from problem solving at multiple scales, which is constitutive of consciousness. Thus, they suggest that a promising way of moving forward is

through dynamical network modeling that explicitly allows for a hierarchy of spatiotemporal levels.

Conclusion

The contributions to this volume differ in their theoretical outlook and in their methodology, which ranges from neuroanatomy, electrophysiology, and functional imaging, through computational analysis and behavioral experimentation, to philosophical analysis. As such, they offer a wide variety of perspectives on the issues about which the authors are in agreement: the dynamical nature of phenomenal experience, and the need to understand and model it as such. Given the profundity and the difficulty of the fundamental questions that arise in consciousness studies, their eventual resolution will require a concerted, interdisciplinary effort on the part of the entire community of researchers. We hope that this book will help generate the right kind of dynamics in this most exciting field of philosophical and scientific endeavor, and look forward to the phenomenal experience of watching this dynamics play out.

Shimon Edelman

Tomer Fekete

Neta Zach

February 2012

Time after time

Temporality in the dynamic brain

Dan Lloyd

Department of Philosophy and Program in Neuroscience, Trinity College,
Connecticut, USA

A useful theory linking dynamical systems to phenomenal experience will be a story thrice told. It will involve some description of phenomenal experience, which should be true. It will also involve some sort of dynamical model (second). But (third) the model will have to be at least plausibly implementable in human beings – that's where the theory becomes useful. Finally, once all three stories are told, they must align. It should be evident to all that the phenomenal story, the dynamical story, and the implementation story are really one story, about one entity, described in three different ways, akin to one story as it might be told in three different languages. A theory of consciousness then is an exercise in translation, somewhat like deciphering the Rosetta Stone. This chapter outlines a possible alignment with respect to a foundational, structural property of experience, namely, time.

1. Introduction: Ubiquitous time

Philosophers in the phenomenological tradition have always stressed the importance of temporality in all experience. For them, our awareness of objects and other people, as well as our awareness of our own psychological states, is not just an awareness of the here and now. Rather, our sensory manifold is embedded in the conscious yet non-sensory context of the past and future, as if the objects of the sensory world each had a time line or history projecting from them. Beginning in 1905, Edmund Husserl (Husserl 1966 (1928)) offered the most elaborate modern account of temporality in perception, an account that has been elaborated but not overthrown in Husserl's successors: Heidegger, Sartre, and Merleau-Ponty, among others. In Husserlian terminology, an "act" of consciousness has three aspects. There is its sensory facet, the information streaming right now from the sensed objects in my immediate perceptual field, which Husserl called the "primal impression." Looking at a cup of coffee, the primal impression comprises sensations and apprehensions of the immediate here and now. And there is the "primary

memory” of the immediate past of every component of awareness. Husserl called this “retention” – that is my awareness of the recent history of my cup of coffee. And there is a future-facing “protention,” which is my immediate anticipation of what will or may happen next. If I grasp and lift the coffee cup, I expect it to have a certain weight and to move freely with my hand. If I turn it over, I expect the coffee to pour out. If the cup turns out to be empty, my surprise reveals the anticipation that was built into my awareness of the cup.

Phenomenologists construe temporality as a structural feature of all states of consciousness, part of the “infrastructure of reality” (Zahavi 1999). Every moment of consciousness is temporal, and every element of conscious awareness has its temporal extensions into the immediate future and immediate past. Protention and retention are folded into the present, at every instance of consciousness. In effect, for Husserl time comprises two dimensions. The vertical structure of protention, primal impression, and retention defines an inner time dimension, and this whole tripartite structure is in continuous flux, a horizontal flow of time in which the three phases of experience are constantly changing. As time flows, protentions are continually drawing near to the primary impression where they will be fulfilled, or not, and primal impressions are continually sliding back into immediate retentional memory. To make things more complicated, the entire tripartite structure is itself a temporal object in one’s awareness. The retention that I experience now of the moment just passed will itself be retained, a retention of a retention. This recursive elaboration is continuous, and extends in anticipation and retention to an indefinite temporal horizon. Over that horizon our temporal awareness is mediated by distinct psychological processes of recollection and explicit anticipation (as would be involved in prediction or planning).

This way of thinking about perception is quite different from the approach taken in mainstream cognitive science. In cognitive science and perceptual psychology, the problem of perception is usually understood as the process of interpreting the immediate sensory field, the primal impression. In David Marr’s classic, *Vision*, for example, visual perception is entirely conceived as a process of recovering three dimensional objects from the flat retinal image (Marr 1982). The perception of the temporal evolution of perceived objects plays no role in this process. Temporality is still largely ignored in contemporary cognitive neuroscience. As in cognitive science overall, neuroscience assumes that perception is driven by the immediate environment, and that the brain is locked or clamped to that environment. So, for example, if I gaze steadily at the coffee cup, my brain enters a steady state of cuppishness, which changes to something different when I look away. But then when I return my gaze to the cup, I return to that same state of cuppishness. Cognitive neuroscience assumes that the brain’s responses to the world are “stationary”, that is, always the same when external conditions are the same and unchanging due to the passage of

time. The assumption that the same perceptual object is correlated with the same state of the brain supports almost every experiment in cognitive neuroscience. In contrast, phenomenal temporality contradicts this assumption of environmental dependency. As I stare fixedly at the cup, my temporal experience of it is continually changing, as I am aware of the ever-changing duration of my looking as a continual elaboration of the retention of my perception of the cup. I'm seeing it for one second, now two, now three, and so forth, and I'm aware of the growing duration of the cup in this location, its history of stability. Eventually I'll achieve the temporal condition of boredom. And if I look back at the cup, my second look is not the same as my first, even if the cup has not changed. I'm aware that I'm seeing the cup *again*, and this is different from seeing for the first time.

Thus cognitive neuroscience omits the consideration of a salient dimension of human cognition and consciousness. Its methods and assumptions are blind to time. But phenomenological temporality is also not without its problems. Husserl describes the flow of temporal experience somewhat like the flow of a river. He acknowledges, however, that temporal experience includes internal distinctions analogous to the experience of spatial objects. For example, interest and attention can highlight some aspects of the temporal field, yielding a foreground and background analogous to figure and ground in visual space. But in general Husserl's writings on time do not offer more than suggestions about how to think about the articulation of subjective time within the foundational structure of protention/primal impression/retention. Phenomenology thus faces a problem that is the converse of that faced by cognitive neuroscience. Against the background of ever-changing time, we nonetheless perceive stable objects and scenes, and recognize their re-occurrence. This too seems like a fundamental aspect of experience.

Both phenomenology and cognitive neuroscience thus leave out important dimensions of human experience, but neither is wholly false. It may be helpful to regard these not as binary opponents but rather as marking two ends of a spectrum of possible models of cognition. Between these antitheses, then, we could look for synthesis, some intermediate analysis that would respect the fact that we do experience a stable world in which objects and situations fall in categories we can recognize over time, while also respecting the fact that time itself generates ceaseless change. If temporality really is a ubiquitous feature of all conscious experience, then the brain must somehow embody temporality.

2. A middle way: Dynamical systems

A dynamical system, like a thrice-divorced heiress, is a system with a past. But unlike persons with baggage, that loaded past is entirely determinative of the

present (and thus the future). To call something a dynamical system, accordingly, is to choose a kind of description rather than to discover some fact of nature, since any system can be dynamically characterized. In this respect the “dynamical systems perspective” is like computationalism, which also describes every system, seen through its lens (Edelman 2008). The DS perspective highlights certain features of systems, and these can be useful. Among these is the idea of a state space, an abstract expanse of many dimensions, one for each variable of interest in the system. Seething vectors condense to wandering points. Geometrical intuitions apply in the state space to a certain extent. For example, distance in its comfortable Euclidean sense can characterize similarity and dissimilarity among dynamical states. But the point analogy is misleading as well. Naked points (or positions) in the physical world are usually uninformative, or in other words points seem not to be capable of encoding complex information. However, in state space, as in the real estate business, location is everything. Magnitude along each dimension of the space is a concrete value encoded in every point.

I look at a coffee cup, and appropriate regions of the brain are excited by its shape, proximity, possibilities, and more. Cognitive neuroscience meets dynamical systems theory when it surmises that each regional activation is a dimension of a brainy state space. The percept of a cup is a point in that space. Phenomenal temporality sets that point in motion. Thus, there is no point corresponding to coffee cups or even this coffee cup. Instead, there is a region in brain space which accommodates the trajectory of the coffee cup. Within that subspace, the cup ten seconds ago is distinguished from the cup now. Furthermore, phenomenology subordinates the Now to retention and protention. Neither of these Not-Nows is part of the current instantaneous environment, so the content of the current point must in some sense encode both the past and future. But how? One way to answer this question is to look and see.

3. Brainspace

The dynamical systems perspective offers some very accommodating metaphors for temporality, but can the trajectories of time be detected in biological brains? The body of this chapter comprises a re-analysis of fMRI data in a particular experiment, a case study affording a look at the methods and data that may inform a scientific neurophenomenology. Can the data of brain science, neuroimaging in particular, be reinterpreted to discover the neural traces of Husserlian temporality? Subjective time should be an aspect of every experience, and thus every brain, provided we use the right interpretive techniques.

Here the target is one neuroimaging study, conducted at the Olin Neuropsychiatric Research Center in Hartford, Connecticut, and I thank the researchers there for their willingness to share their many gigabytes of data.¹ In this experiment, participants drive a simulated automobile through a virtual landscape, controlling their route with a steering wheel, and brake and gas pedals. The experiment is divided into three runs, and each run has the same structure of three parts: First, subjects see a black screen with a plus sign, a “fixation cross,” in the middle, for 30 seconds. Then the virtual reality simulation begins, and subjects drive through the simulated landscape, for 90 seconds. Then, subjects move through the simulated landscape again, but it is as if they are passengers in the automobile, merely observing but not driving. This lasts for 60 seconds. This same sequence of Fixate, Drive, and Observe repeats three times without intermission, and is followed with a final thirty seconds of fixation. The whole session, then, comprises this sequence:

Fixate – Drive – Observe – Fixate – Drive – Observe – Fixate – Drive – Observe – Fixate

With three runs, the experiment lasts about nine minutes. While the subjects are driving, a number of behavioral variables are recorded. This includes the current position of the car in the virtual landscape, steering, accelerating, and braking, and others. Each session yields 375 whole-brain images, collected every 1.5 seconds. Although no one would confuse the experience in the scanner with actual driving, it is nonetheless a compelling and immersing perceptual and behavioral task. Complex and open-ended, the task comes closer to simulating a natural interaction with the world than the host of button-pushing stimulus-response experiments designed to probe specific narrowly defined facets of perception or cognition. In addition, each session in the scanner comprises three very different subjective states, Fixation, Driving, and Observing, which interweave and repeat.

In their phenomenology the three conditions are certainly diverse: One of them requires action, while two are passive. One presents a nearly blank screen, while two are visually rich and dynamic. Facing this experiment, the standard neuroimaging approach would seek correlations between activity in specific regions

1. Publications from this study include: Carvalho, K.N., G.D. Pearlson et al. (2006). Simulated driving and brain imaging: Combining behavior, brain activity, and virtual reality. *CNS Spectr* 11(1): 52–62. Calhoun, V.D., J.J. Pekar et al. (2004). Alcohol intoxication effects on simulated driving: Exploring alcohol-dose effects on brain activation using functional MRI. *Neuropsychopharmacology* 29(11): 2097–2017, Calhoun, V.D., K. Carvalho et al. (2005). Using virtual reality to study alcohol intoxication effects on the neural correlates of simulated driving. *Appl Psychophysiol Biofeedback* 30(3): 285–306.

of the brain and particular configurations of stimuli and responses, ultimately to explain the behaviors elicited by the experiment as the joint production of functionally specific brain regions. The phenomenology of temporality, however, demands something more than the parceling of brain states into bins to match the experimental task. For starters, the repeating experimental conditions cannot be combined in an analysis of temporality. The first, second, and third runs retain their distinctness in experience. The time line of the experiment must at least be adumbrated to this extent:

Fixate(1) – Drive(1) – Observe(1) – Fixate(2) – Drive(2) – Observe(2) – Fixate(3)
– Drive(3) – Observe(3) – Fixate(4)

Moreover, within this articulation temporality demands a continual awareness of passing time. Thus, each moment of the experiment is further articulated by temporality itself, the awareness of elapsed or anticipated durations, embedded in consciousness as a structural feature of each moment of awareness:

Fixate(1.1) – Fixate(1.2) – Fixate(1.3) – Fixate(1.4) – ...

At this point the divergence of cognitive neuroscience and phenomenology is sharply drawn. Phenomenology says to neuroscience, show me how Fixate(1.1) is different from Fixate(1.2) and (1.3), etc. and show me how these are different from Drive(1.1), and furthermore how Drive(1) differs from Drive(2). The neuroscientist says back to the phenomenologist, show me how Fixate(1.1), (1.2) and (1.3), are the same, and likewise how the three drive epochs share a common functional basis, and how the three experimental conditions differ.

Wrestling with these questions conversationally leads to dead ends. Thinking of state spaces, on the other hand, leads directly to a reconciliation. Two states are similar if they are nearby in state space, and different to the extent that they are mutually distant. This sliding distinction allows sameness and difference to cohabit the same data. In the present case, we imagine a state space containing three broad territories, mapping the three experimental conditions. But within each territory, we anticipate a spatial subdivision corresponding to the three instances of each condition. So, Drive(1) and Drive(2) can occupy their own counties within the territory shared by the driving condition. Furthermore, the fine-grained temporal distinctions, the ticking of the subjective clock, can be captured as well, as baby-steps within the overall space of each condition in the experiment.

State spaces are capacious, accommodating as many dimensions as there are variables of interest. In the brain, the variable of interest is commonly assumed to be the neuron (or perhaps the synapse) – these have been satisfying basic entities in small scale explanations of signal detection, learning, and behavioral coordination, not to mention their fit in biological explanation overall. The state

space of neurons in a typical human brain is high dimensional indeed. We can't track that many variables and even if we could, it's not clear what would be gained – this is the “curse of dimensionality” (Fekete & Edleman 2011). A brain equation of one hundred billion variables would be of slight practical use. Accordingly, with high dimensions comes the impulse to simplify, that is, to combine or eliminate dimensions to build a useful subspace.

With the present experiment, data reduction occurs twice over. First, fMRI is both indirect and blurry in its detection of neurons in action. It's indirect, insofar as it tracks the blood oxygen level dependent (BOLD) response to neuronal activity. It's blurry, in that every voxel is an amalgam of millions of neurons smeared over 1.5 seconds (Logothetis 2002). Despite that formidable funneling, each image in the driving dataset comprises 150,000 voxels, still an unwieldy multitude. The second wave of simplification, then, is Independent Component Analysis (ICA) (Calhoun, Adali et al. 2002). The “components” discovered by ICA are ensembles of voxels that activate and deactivate in unison, “temporally coherent networks” of correlated activity. These voxel teams can be localized but are often distributed across two or more regions of the brain. Each component sings with one voice, but is uncorrelated with the other components. Around twenty components can retain nearly 90% of the variance in most brain data sets.

So billions of neurons appear to the researcher as a temporally and spatially blurred vector of just twenty elements. That is, the state space in this analysis has a mere twenty dimensions. At this point, a sensible person would doubt the possibility of a fine grained analysis of the sort needed to identify the trajectory of time. Nonetheless, the spark of curiosity invites some exploratory data analysis. To begin, we can represent the brain space in even fewer dimensions, using Multi-Dimensional Scaling (MDS), as in Figure 1. Here, a two dimensional map preserves the distances between the points in 20-d brain space, for ten subjects in the driving experiment. Each time point is mapped, and coded for the big distinction between Fixate, Drive, and Observe conditions.² Inspection reveals a promising regional distinction among the three conditions, which has been crudely marked – one could draw the boundaries many different ways. The map conforms to a very basic intuition about this experiment, namely, that brains are in different global states during the three conditions. Cognitive neuroscience and phenomenology converge to this extent at least.

2. MDS was implemented using the `mdscale` function in MATLAB. The analysis was based on the 20-element vectors of ICA magnitude at each of the 375 time points in the experiment. The ten subjects were concatenated, so that each time point was represented as a 200-element vector. From this Euclidean interpoint distances were calculated, and non-metric MDS performed.

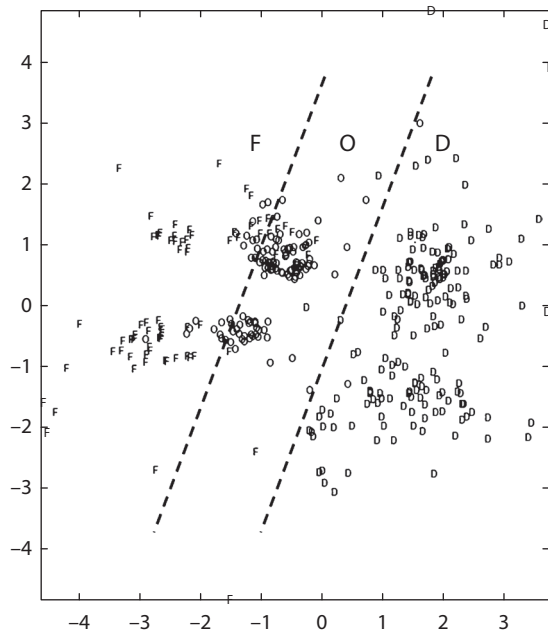


Figure 1. State space of 375 images of the driving experiment, combined across ten subjects. Twenty dimensions have been reduced to two with multidimensional scaling. F: Fixate conditions; O: Observe conditions; D: Drive conditions. For the most part, similar conditions are similarly located in brainspace, as suggested by the dashed dividers

Although the MDS map in Figure 1 is helpful for visualizing brain space, it is also quite inaccurate, since high dimensional inter-point distances must be squeezed and stretched to force 375 points onto a flat surface. “Stress,” the term for map distortion in MDS, is high. Cluster analysis, another exploratory method, is not so stressful. We can confirm our map intuitions by clustering the nearest points in the high-dimensional space. Carving brain space into three clusters shows that the distinctions are indeed sharp. Figure 2 is a cluster diagram of a slightly non-traditional format. Clusters are distinguished along the y-axis, and the time points of the experiment pass from left to right. Plusses stand for each point in state space, each assigned to its cluster.

Cluster analysis is more than a convenience for data visualization, however. Cluster membership and boundaries demarcate the capacities of a system for making distinctions. Fekete and Edeleman write:

The basic means of realizing conceptual structure is *clustering* of activity: a representational system embodies concepts by parceling the world (or rather experience) into categories through the discernments or distinctions that it

induces over it. As it gives rise to experience... activity should possess no more and no less detail than that found in the corresponding experience.... [A]ctivity space must divide itself intrinsically into compartments, structured by the requisite within- and between-concept similarity relations.

(Fekete & Edelman 2011, p. 5)

Figure 2 confirms that the gross distinction between conditions in this experiment are represented in the global brain space. The two stages of fMRI observation and ICA data simplification have unwoven the curse of dimensionality without losing the conceptual structure we expect of this cognitive system.

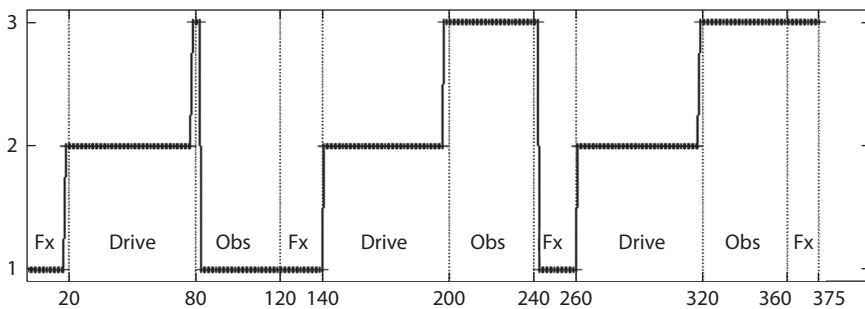


Figure 2. Brainspace in the driving experiment divided into three clusters. Each image is represented by + at its time point, assigned to its cluster. (Clusters are numbered (arbitrarily) on the y axis.) Clusters are based on Euclidean distance and partitioned with Ward's method. Each condition is assigned to its own cluster, with the exception of the first Observe and last Fixate conditions

Brain state, cognitive science, and phenomenology converge on the common sense carving of the state space of the experiment. Images from the Drive condition are clustered with 99% accuracy. Overall, the clustering is 83% correct.

Once we go beyond three clusters, however, we can begin to explore brain-based categories, to probe whether further divisions are driven by time. By constraining the analysis to n clusters, we examine the dynamical brain at different levels of granularity (Fekete & Edleman 2011), depending on n . Temporality is displayed if at some granularity temporally adjacent time-points are clustered together. Figure 3 displays cluster anatomy at 5, 8, 12, and 20 clusters, successively zooming in on the fine structure of brainspace. Figure 3A, with five clusters, already displays the temporal separation of the first occurrences of the three experimental conditions. Overall, the state space segregates into early, short segments and later, longer segments. This repeats at different scales. The Drive epochs also resolve into repeating short-long units (panel B, with eight clusters). Twelve clusters separate all three

conditions into separate spaces (panel C). Finally, with twenty clusters available, the Observe blocks separate into their temporal parts, while the Drive epochs further subdivide. Overall, the largest territories are categorical, reflecting the distinctions between the experimental conditions, but these resolve into temporally contiguous subgroups. Time appears to be the modifier of all three conditions. Finally, Figure 4 curves brainspace into 60 territories.

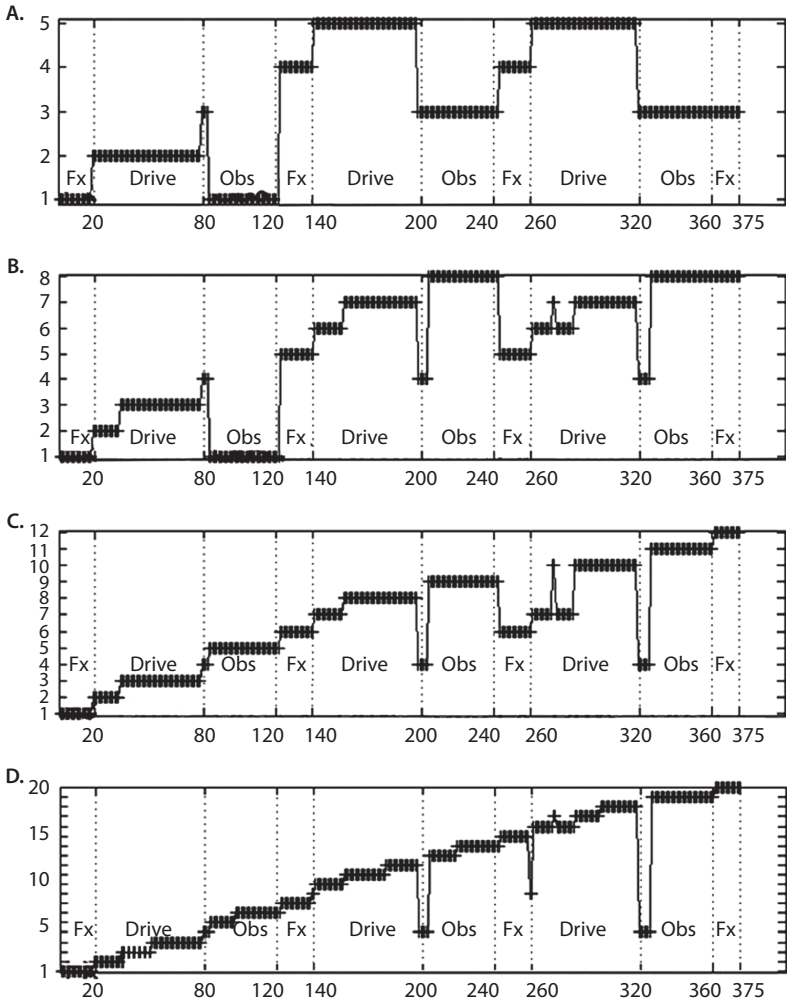


Figure 3. Cluster analyses at different levels of subdivision. A. Five clusters (numbered arbitrarily on the y axis) show separation of early and later epochs along with segregation by conditions. B. Eight clusters show within-condition subdivisions, also based on time. C. Twelve clusters almost completely separate same conditions by time. D. Twenty clusters show groups almost entirely segregated by temporal sequence

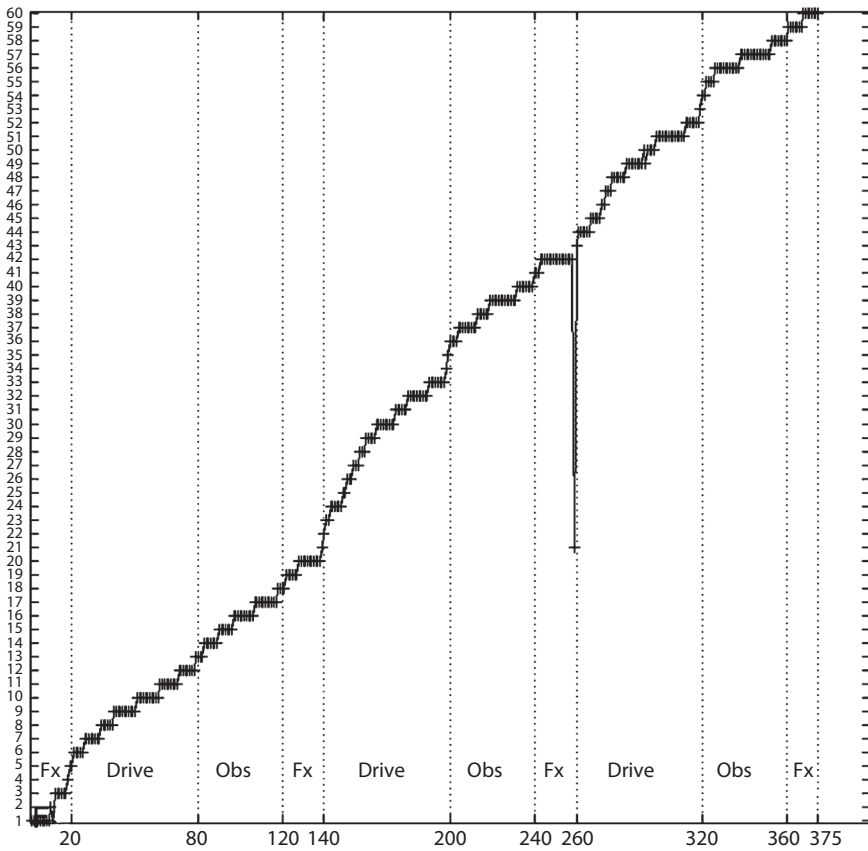


Figure 4. Cluster analysis of brainspace trajectory across sixty clusters. Temporal divisions dominate. At this grain, velocity through brainspace becomes visible as time spent within each cluster. Cluster jumps are more frequent at condition boundaries, for example

At this level, aspects of the flow of time become visible. Note the boundaries of the Drive conditions, at which the brain moves rapidly through distinct states. The overall “velocity,” or the number of cluster jumps over time, varies with the different conditions. Brain activity seems to change most quickly at the boundaries of conditions. Within conditions, Drive and Observe clusters pass more quickly (compared to the expected cluster size with a uniform distribution). Here we can see the rapid jumping from state to state in the eventful Drive sequences. In contrast, time drags during the boring Fixation conditions. But looking at these episodes retrospectively changes how they seem. The many events of the Drive periods might make those sections seem longer in retrospect, in contrast to the Fixation sections, which in retrospect might seem shorter than their actual duration.

Considered together, then, as the view through brainspace zooms, both traditional phenomenology and the assumptions of cognitive neuroscience are accommodated. At the coarsest level, the three very different conditions of the experiment push the brain into three different distributed states, just as cognitive science sees it. But within each of those territories, brains wander about. Nonetheless, the wanderings keep time points distinct and ordered, just as the phenomenologists insist.

4. Temporality now!

Cluster analysis of brain space in this experiment reveals different facets of brain function at different magnifications. At a coarse grain, the brains analyzed here are categorical engines, switching from one global state to another as the task demands. But at a finer grain, within each global category, brains are time-keepers – notwithstanding the fact that subjects in this experiment were not required to notice the passing of time. Initially, therefore, we observe a correlate to structural, ubiquitous time.

Clocks also change with passing time, but clocks show the time without knowing the time. Clock time is not temporal awareness for the clock. Apparently, temporal information is encoded in state space, but this is not yet to establish that this information is encoded in the brain itself. What more is needed? Phenomenological temporality embodies a three way distinction between protention, primal impression, and retention. These are three co-occurring aspects of every state of consciousness. Accordingly, their brainy implementation must not only tag the current time, but encode past and future as well. With this requirement, temporality pulls the brain still farther from the stationary, task-clamped organ described in standard cognitive neuroscience. Each brain state not only encodes its current location in a temporal landscape, but also its past and future positions. It not only has a trajectory but somehow encodes that trajectory as well.

Trajectory through brainspace is at the heart of the proposed computational theory of experience of Fekete and Edelman (2011), and motivated in part by the phenomenology which is also central to this chapter:

A model in which qualia are equated to instantaneous independent states would be hard pressed to explain various mundane cognitive phenomena, not the least of which is the experience of time. (p. 9)

Indeed, but the phenomenology does not merely imply that dynamical brain configurations evolve along a path over time but that instantaneous states

themselves represent this trajectory. That's the Husserlian bottom line, the vertical dimension of temporality that shadows the horizontal flow of time. This instantaneous synchronic representation of trajectory (which may be both gappy and inaccurate) is distinct from the diachronic trajectory itself.

What empirical evidence might be enlisted to confirm this temporal extension? Here we take an indirect route, inspired by work in machine learning and Support Vector Machines (SVMs) in particular (Cox & Savoy 2003; LaConte, Strother et al. 2005). SVMs learn to map a set of input patterns to a corresponding set of output patterns. Their learning is confirmed when an additional, unlearned, set of input patterns is correctly associated with target outputs. Successful association on the novel inputs implies that they must themselves encode information in their distributed activity to enable the SVM to project the desired target. This is the consideration that we will exploit to search for encoded time in patterns of brain activity.

Specifically, we can use SVMs to probe whether a pattern of brain activity at time t_n can predict activity at t_{n+i} where i ranges into the near future, and likewise retrodict/recover past states. In the driving experiment, can SVMs take the brain state at (for example) 22 seconds into the experiment, and correctly anticipate experimental conditions at 23, 24, 25... seconds, as well as correctly reconstruct past patterns at 21, 20, 19... seconds. As in other SVM paradigms, here we test accuracy with a "leave one out" cross validation strategy. During training, one vector of activations is omitted from the training set, and reserved for later testing. The accuracy of SVM training is then measured by the accuracy of classification of the omitted vector. This is repeated for all inputs to provide a general measure of SVM learning.

Figure 5 displays the landscape of protention and retention as revealed through the SVM train and test paradigm, showing percent accuracy of SVM time projection from each of 375 test points. The time unit is 1.5 seconds, the time taken for each image in the fMRI series. The figure thus summarizes the temporal landscape looking 15 seconds forward and back from each time point. The left half of the graph represents retentional information recovered from brain images at different lags from the current time, showing that accuracy is greater in recovering the immediate past, compared to the immediate future.³

3. SVMs are binary classifiers, so the experimental data set was decomposed into three binary sets, one for each of the three experimental conditions, Fixate, Drive, and Observe. So, for example, Fixate conditions would be encoded as +1, and all others -1, etc. Furthermore, separate SVMs were trained and tested for each time offset, moving the +1/-1 boundary from 15 seconds behind the actual transition to 15 second ahead. The resulting scores are aggregates from these various sub-conditions.

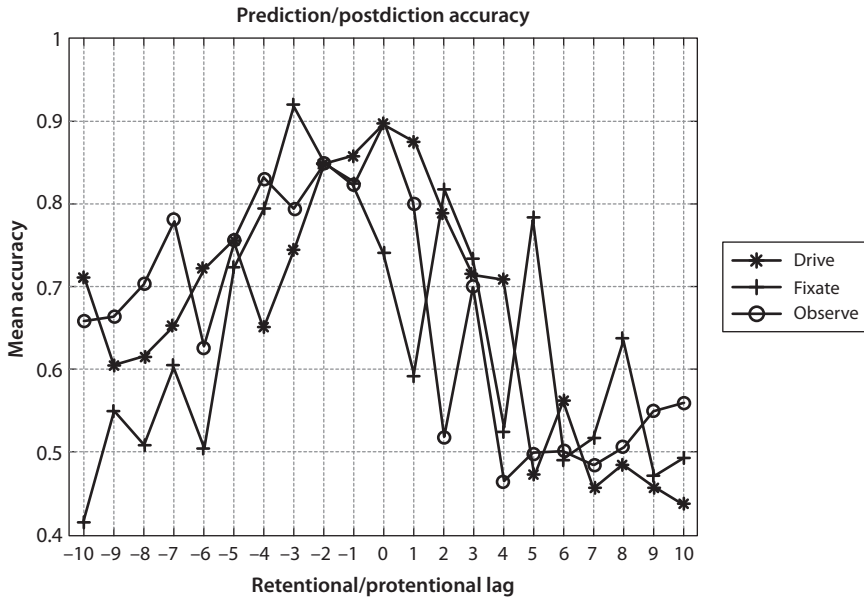


Figure 5. Embedded temporal information in each image in the drive experiment, mean of 375 times points in ten subjects. Accuracy is measured with a “leave one out” cross-validation strategy, measured on each image when it has been omitted from the training set. Because support vector machines are constrained to sort targets into binary categories, each experimental condition was tested separately. The graph implies that from each image, in general, projections are more than 50% accurate about six seconds into the future (4 time points on the graphic), and retentional information is accurate nine or more seconds into the past. (Chance performance is 50%.) At each time point, the brain appears to represent a temporal window of about fifteen seconds

5. Countdown to implementation

The evidence presented so far suggests that brains encode retentional and protentional information for several seconds and possibly longer, and that traces of this encoding are captured in fMRI image series. The impulse to localize this information is strong. But rather than follow this path, a final empirical foray explores a more general question about implementation: Is temporal information localized at all? At one extreme, one region of the brain may be the time-giver, a dedicated clock that can somehow be read by the SVM-fMRI combination. At the other extreme, time could be embedded in configurations of states, a distributed pattern of activation whose changes represent the evolving temporal landscape of this experiment. Phenomenology suggests that temporality is in every experience, suggesting a more distributed implementation. This is the suggestion to be tested here.

The test strategy is as follows: We successively remove each of the twenty independent components, and test the remaining nineteen using the methods of the previous section (Oppel & Winther 2000). Performance degrades following some of these virtual lesions. The component whose absence most impairs performance is thus the component of greatest import in temporal representation. If performance dropped to chance levels, then we would have identified a component necessary to temporal encoding. But if temporal representation persists, even if somewhat impaired, we can repeat the process, now with the reduced set of nineteen components. That is, we can successively remove components one-by-one and test the remaining eighteen to discover which component contributes the most to pattern learning performance. This will be the second-most important component overall, and it too is set aside as the analysis continues, now lesioning each of the remaining eighteen components. And so forth, leading eventually to Figure 6, plotting the mean performance as the number of components is

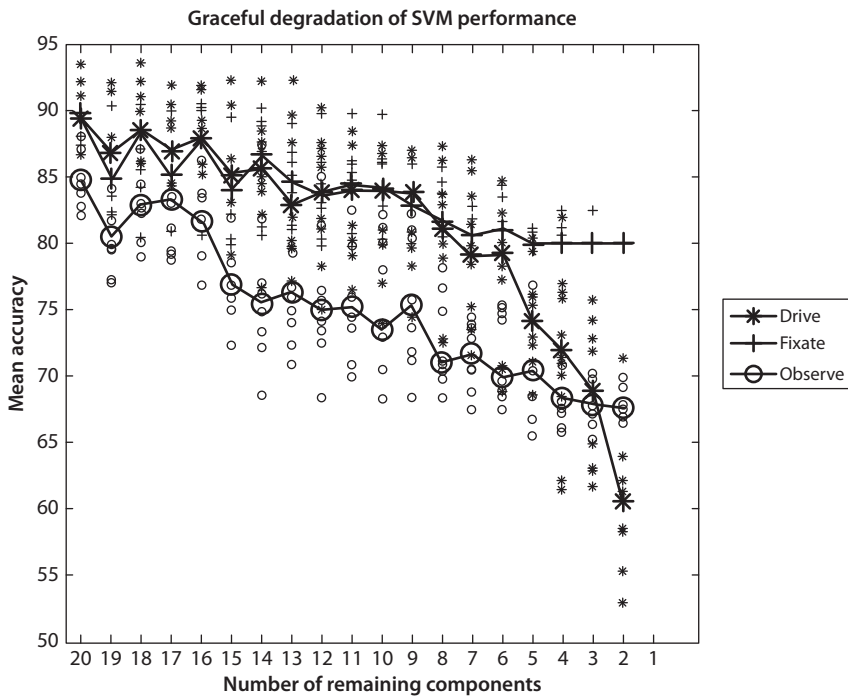


Figure 6. The effects of successive removal of the most effective components. For each subject, accuracy is tested with one component removed to determine the most important component (i.e. the components whose removal has the greatest impact on accuracy). That component is removed, and the process repeated on the reduced component set. Performance falls steadily toward chance levels in most subjects. Large symbols represent the median accuracy of the SVMs for 10 subjects. Small markers represent accuracy individually

decremented, removing the most important component with each decrement. The figure shows a graceful decline as approximately ten components are removed. Beyond ten, overall performance falls off more rapidly. Analysis using this “leave one out” strategy thus implies that temporal information is broadly distributed across regions of the brain. The encoding of protention and retention within the 30 second window considered here seems to depend on a varying pattern of activity rather than a localized clock.

6. Crossing the explanatory gap

The science of consciousness begins with correspondences between items in awareness and items in the brain, just as deciphering a text in an unknown language might begin by identifying synonyms between the known and the unknown lexicons. A large part of the study of a new language is just this, and likewise a great deal of consciousness research is a search for “neural correlates of consciousness.” As the correlates accumulate the ties of brain to phenomenology diversify and skepticism about materialism becomes increasingly untenable. But a pile of correlates do not a theory of consciousness make, just as a lexicon is only one aspect of a language. A *theory* of consciousness is an edifice that shows how all the pieces fit together, comparable to an explicit grammar for a language. If the particulars of the neural correlates of consciousness are the first-order observations in the science of consciousness, then we must also look toward higher-order relationships as well.

Dynamical systems offer a useful framework for theorizing. State spaces afford a way to conceptualize higher order relationships that balances multivariate complexity against the limits of human imagination. Here the dynamical systems analogy structured an exploration of the higher-order neurophenomenology of time. In this one case study (with a mere ten subjects), we observed brainspace relationships that respected first-order distinctions (between the conditions of the experiment) while at the same time reflecting temporal distinctions between repetitions of the experimental conditions and within same-condition blocks. At this point, however, the analogy with translation falters. It’s not enough to show that the dynamical states of the brain reflect the trajectories of conscious awareness. Consciousness is instantaneously temporal: each state of awareness includes an awareness of whence and whither, retention and protention. This seems to be more than just its dynamical causal tendencies, but rather an active presentation of temporal extension. Higher-order constraints govern the system and also appear in some form within states of consciousness.

To locate a synchronous representation of diachronic time required a more elaborate probe of the same brainstate data. Here the approach was indirect, using machine learning to seek the patterns across the brain that might embody the forward- and backward- gaze of temporal awareness. It was argued that if a support vector machine could successfully extract the experimental conditions both before and after the current time point, then the temporal information must already be in the functional brain image at that time point. Success in this learning paradigm leads naturally to an interest in its “neural correlate.” In this case, we want to know which areas of the brain support the representation of time. The current chapter stops short of that answer, but does present evidence that broadly distributed patterns of activity are involved in the representation of temporal information.

Suppose, then, that further confirmations follow, based in different experiments and new methods. To what extent do reports like these inform a genuine theory of consciousness? Two general answers suggest themselves. First, it is essential that theories of consciousness address higher-order features of experience. Time is an example of a higher-order property, a structural aspect of all states of awareness. Temporality infuses everything else, and so every first-order neural correlate is also the vehicle of the higher-order structural awareness of time. Merleau-Ponty declared that “we are time,” affirming the centrality of temporality to consciousness (Merleau-Ponty 1962). Temporal relationships provide a basis for many seemingly non-temporal properties. Space, for example, could be construed as a set of possible physical trajectories for objects of perception, such trajectories being represented in retention and protention. Objects are equally temporal, as Husserl himself described at length in 1907 (Husserl & Rojcewicz 1997).

More important, self-consciousness could be reanalyzed as a temporal trajectory as well. Self-consciousness and its problems are frequently taken to be the problem of consciousness overall. In my opinion, this is a mistake. Although all states of awareness are inflected by a point of view, they do not require a further reflective awareness of this point of view. In the phenomenological tradition, we encounter a distinction between consciousness “in itself” and consciousness “for itself” (As Sartre puts it: *en soi* and *pour soi*. (Sartre 1958)) Sartre, in accord with Husserl and Heidegger, affirms that much of conscious life is “in itself,” that is, unaccompanied with self-awareness (for itself). Promoting self-awareness into an essential condition for experience leads to either an infinite regress or the implication that states of self-awareness (higher order thoughts) are not themselves conscious. On a more practical level Fekete and Edelman (2011) have argued that a synchronous awareness of the current state of the brain by that self-same brain is computationally infeasible. Temporality allows us to reconfigure self-awareness as a reflective access to retentional information about one’s own prior states (both

percepts and concepts), a modest and tractable self-awareness that better fits with the real phenomenology of self.

Thus, many aspects of conscious life can be reconstrued as an awareness of trajectory over time. A theory of consciousness that includes temporality is a worthy advance from the accumulation of first-order NCCs. But this brings us to the brink of another canyon of mystery: How does it work? The analyses in this chapter have been silent on the actual mechanisms of temporal awareness. Recurrent networks (“reentrant processing”) appears repeatedly in proposed mechanisms of consciousness (Edelman 1993; Di Lollo, Enns et al. 2000; Clavagnier, Falchier et al. 2004; Lloyd 2004; Augustenborg 2010; Miller 2011). Such networks have the capacity for temporal representation. Indeed, they require it if they govern behaviors that extend over time. Unfortunately, functional MRI may be too coarse an instrument to probe these dynamics.

Which is more complex, the brain or consciousness? Both are vast, and it is no surprise that their union in the science of consciousness will be a long story indeed. But every story has its beginning. Among the proliferating beginnings by scientists and philosophers, we may find one or more with promise. Once upon a time....

Acknowledgements

Thanks to: Vince Calhoun and Godfrey Pearlson, who provided the study data sets, and to the Trinity Faculty Research Committee for ongoing research support. Independent Component Analysis was implemented with the ICA Toolbox (Calhoun, Adali et al. 2002). Support Vector Machine analysis was implemented with the OSU-SVM Toolbox (Ma, Zhao et al. 2002).

References

- Augustenborg, C.C. (2010). The endogenous feedback network: A new approach to the comprehensive study of consciousness. *Conscious Cogn* 19(2), 547–579.
- Calhoun, V.D., Adali, T. et al. (2002). Independent component analysis of fMRI data in the complex domain. *Magn Reson Med* 48(1), 180–192.
- Calhoun, V.D., Carvalho, K. et al. (2005). Using virtual reality to study alcohol intoxication effects on the neural correlates of simulated driving. *Appl Psychophysiol Biofeedback* 30(3), 285–306.
- Calhoun, V.D., Pekar, J.J. et al. (2004). Alcohol intoxication effects on simulated driving: Exploring alcohol-dose effects on brain activation using functional MRI. *Neuropsychopharmacology* 29(11), 2097–2017.
- Carvalho, K.N., Pearlson, G.D. et al. (2006). Simulated driving and brain imaging: Combining behavior, brain activity, and virtual reality. *CNS Spectr* 11(1), 52–62.

- Clavagnier, S., Falchier, A. et al. (2004). Long-distance feedback projections to area V1: Implications for multisensory integration, spatial awareness, and visual consciousness. *Cogn Affect Behav Neurosci* 4(2): 117–126.
- Cox, D.D. & Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) “brain reading”: Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19(2 Pt 1), 261–270.
- Di Lollo, V., Enns, J.T. et al. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *J Exp Psychol Gen* 129(4), 481–507.
- Edelman, G.M. (1993). Neural Darwinism: Selection and reentrant signaling in higher brain function. *Neuron* 10(2), 115–125.
- Edelman, S. (2008). *Computing the mind: How the mind really works*. New York: Oxford University Press.
- Fekete, T. & S. Edelman (2011). Towards a computational theory of experience. *Consciousness and Cognition* 20, 807–827.
- Husserl, E. (1966/1928). *Zur Phänomenologie des inneren Zeitbewusstseins (Phenomenology of Inner Time Consciousness)*. The Hague: Martinus Nijhoff.
- Husserl, E. & Rojcewicz, R. (1997). *Thing and space: Lectures of 1907*. Dordrecht; Boston, Kluwer Academic Publishers.
- LaConte, S., Strother, S. et al. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26(2), 317–329.
- Lloyd, D.E. (2004). *Radiant cool: A novel theory of consciousness*. Cambridge, Mass.: MIT Press.
- Logothetis, N.K. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philos Trans R Soc Lond B Biol Sci* 357(1424): 1003–1037.
- Ma, J., Zhao, Y. et al. (2002). OSU-SVM Toolbox for Matlab.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.
- Merleau-Ponty, M. (1962). *Phenomenology of Perception*. London: Routledge and Paul.
- Miller, G. (2011). Neuroscience. Feedback from frontal cortex may be a signature of consciousness. *Science* 332(6031), 779.
- Oppel, M. & Winther, O. (2000). Gaussian processes and SVM: Mean field results and leave-one-out. *Advances in large margin classifiers*. Cambridge, MA: MIT Press, 43–65.
- Sartre, J.-P. (1958). *Being and nothingness*. London, Philosophical Library.
- Zahavi, D. (1999). *Self-awareness and alterity*. Evanston, IL, Northwestern University Press.

Neuronal reflections and subjective awareness

Rafael Malach

Department of Neurobiology, Weizmann Institute of Science

For a neuroscientist working under the assumption of a complete correspondence between mind and brain, conscious awareness poses a profound mystery. It is a unique phenomenon in which a state of a physical system – i.e. a distributed pattern of neuronal activity is inexplicably transformed into an absolutely private and internal mental experience. Thus, the relevant point of view that needs to be taken when considering neuronal mechanisms underlying phenomenal experience must also be intrinsic – that of the brain's neurons themselves. In contrast, the perspective taken by most neuroscience research examines neuronal activity from an external observer perspective. Here I will conjecture that the intrinsic perspective is implemented in the brain through rapid and recurrent neuronal activity – local neuronal “ignitions”. In this dynamics, the information about the state of a local neuronal assembly is distributed back to the neurons that form the assembly through recurrent activations. A conscious percept emerges when, through these neuronal reflections, the ambiguity inherent in the activity of isolated neurons is converted into a unique and meaningful assembly state. The rapid distribution of assembly information necessitates high firing rates, sustained activity, and dense local connectivity. All these conditions fit nicely with recent experimental findings. The fourth consequence – the founding of conscious awareness on local reverberatory activity – is still highly controversial – and should be viewed, at this stage, as a prediction of the local “ignition” hypothesis.

1. Introduction – the optimistic outlook

It is a self-evident fact that our subjective, phenomenal state, e.g. the image we have when marveling at the starry night, is the most direct and accessible of all occurrences we can conceive of. Thus, one would expect that this universe of sensations will be high on the priority list of neuroscientists to explore and explain. Yet, the contrary is true – the scientific exploration of phenomenal experience has been extremely hesitant and plagued by controversies. Thus, it is still debated whether subjective experience is at all amenable to proper scientific research. The reason is of course the very unique nature of subjective experience – which drastically differs from all other natural phenomena. Particularly problematic to

scientific research is the private nature of subjective experience which is inaccessible to direct measurements or objective observation.

Furthermore, unlike all other physical phenomena – the link of subjective experience to the physical world is deeply unintuitive and in fact incomprehensible. It is an enigmatic process by which a physical system – i.e. a collection of nerve cells in an individual's brain – is somehow capable of “becoming aware” – i.e. transforming a distributed pattern of neuronal activity into an absolutely private and unitary phenomenal experience. Note that this statement makes two implicit assumptions, which should be stated upfront. First, it is assumed that phenomenal experience is indeed linked to the brain – the philosophical stance I will hold is that of monism – as formulated already by Spinoza’ – i.e. that subjective experience and the physical state of the brain are two different perspectives of one and the same phenomenon. The second assumption is that there is a one to one mapping between contents of phenomenal experiences and physical states of the brain. Discussing alternatives to these basic assumptions will be beyond the scope of this chapter.

It has been extensively debated whether the unique nature of phenomenal experience may actually preclude the construction of a scientific theory of conscious awareness – which is after all the goal of any scientific endeavor. Here I will adopt a more optimistic outlook and will argue that substantial progress in the scientific search for the neuronal basis of subjective experience is feasible. However it necessitates a major shift in our point of view when considering models of phenomenal awareness. If we accept that subjective experience entails some internal link to the state of the brain's neurons – then as scientists trying to understand this process, we must forego our natural perspective as experimentalists – that of external observers collecting information about the external reality, and instead adopt an “intrinsic” perspective. Such perspective should examine what information about the state of the brain is actually available to its constituent neurons on their own – unaided by the knowledge we experimentalists gain in the course of our observations.

As an example for subjective phenomenal experience I will use conscious visual perception – since my research has mainly focused on this type of experience. But it should be possible to generalize the considerations derived from this research to the entire set of phenomenal subjective experiences.

2. The creative nature of visual perception

Starting with Helmholtz and the Gestalt Psychologists – our understanding of the behavioral aspects of visual perception has shown a tremendous progress leading to the establishment of visual psychophysics as a rigorous and rich field of active

research. Here I will focus only on few general principles that this research has led to and which are directly relevant to the argument I am trying to develop. One such fundamental principle is the understanding that the content of a conscious visual percept (which I will operationally define as any visual experience that a healthy individual can both introspectively and objectively report about) can often go far beyond the immediate optical or retinal information provided by the eyes. Two classic examples that illustrate this point are depicted in Figure 1. In the phenomenon of illusory contours (e.g. the Kanizsa figure, panel A) and in the case of “impossible figures” – panel B. In both cases the phenomenal percept actually *contradicts* the physical objective information. The physical world and the retinal image consist of isolated wedges in the case of the Kanizsa triangle, yet the phenomenal percept contradicts this objective information and generates instead a subjective experience of a triangle which does not exist in reality. While in the case of the illusory triangle one may still invoke prior experience or memory with such triangles as the basis of the “illusion” – in the case of the impossible figure even this can not be true – since such objects can not possibly exist in the real world – and hence, surely were not experienced by the observer prior to seeing them.

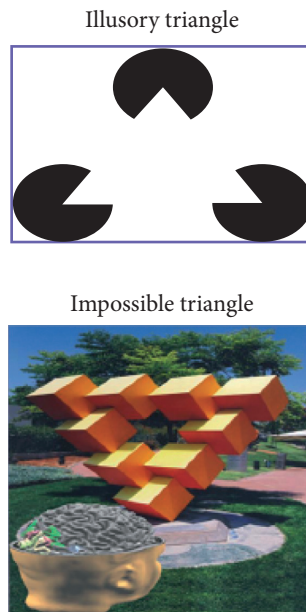


Figure 1. Two visual illusions illustrating the discrepancy between visual perception and objective reality. The illusory triangle is visible despite the physical absence of edges. The “impossible” triangle is easily perceived as a triangle despite the fact that such objects can not possibly exist in the physical world

These are of course just two examples out of a large set of striking illustrations – all pointing to an unavoidable conclusion – the visual percept can not be viewed as some kind of transformation, processing or “representation” of the physical visual information. A straightforward consequence of this is that the information that underlies the creation of a phenomenal visual image must be available internally – within the brain itself. The most striking illustration of this fact is in the spontaneous emergence of vivid visual images during dreaming as well as following electrical stimulation in patients, in visual hallucinations or upon trans-magnetic stimulation (TMS). In all these cases, a vivid visual percept is generated in the complete absence of any optical stimulation or an external visual target. This is not meant to imply that the external world is irrelevant to perception – to the contrary – clearly a large part of the internal information underlying the phenomenal image reflects regularities and adaptive values of the external physical world. The point is that during a moment of a perceptual awareness, when we see a vivid visual image, the information pertaining to this image is derived mainly from the brain itself. Thus, it is in the brain, not in the outside world, that we should search for the source of information allowing the moment by moment creation of visual percepts.

3. A colossal library

A phenomenal visual percept can not therefore be viewed as some kind of transformation of optical information derived from the external world, but is rather derived from information stored internally in the visual system itself.

The significance of this conclusion becomes evident when we consider the sheer number of different visual percepts that a single human visual system is capable of generating. Taking just the case of faces, a person is quite likely capable of differentiating between most faces belonging to one’s own race – likely millions of different faces. Considering the fact that each of these individual faces can appear in different colors, sizes, contrasts, expressions, ages etc. – it becomes clear that our capacity to generate differentiable visual contents runs in the many trillions of different images.

It is important to clarify that we are not dealing here with invariant visual recognition or memory – where the capacity is likely far more limited. The challenge for a theory of phenomenal visual experience concerns the staggering amount of potential visual images – i.e. we need to explain how brain states can be constructed so that such an enormous set of subjective percepts can be differentiated. It is obvious but needs stressing, that when it comes to estimating our capacity to generate conscious visual percepts – any slight consciously perceived

changes – such as color changes of the same person – must give rise to different brain states. With this in mind, it is clear that the image generating capacity of the brain is truly colossal. So how are neurons capable of generating such a massive number of differentiable visual percepts?

4. Building templates – hierarchically

In trying to come up with a solution, I will focus on the primate cortex – since it is in primates that we assume that the link between behavioral report and subjective experience is most reliable. Two points should be noted before we embark on this analysis. First, the fact that it may be more convenient to study primate vision does not imply that other animals have a reduced level of phenomenal visual experience. Second, even in humans, the ultimate inaccessibility of external observation to the subjective content of experience should be acknowledged – i.e. we can never be completely certain whether the report we receive truly reflects the phenomenal experience. Nevertheless, as I will try to illustrate below, some conceptual advances can be made with regards the relationship between subjective states and brain states.

Examining how the processing of visual information begins in the primate cortex, the first principle that becomes apparent is the hierarchical nature of the processing. Thus, converging evidence supports the notion that as we move anteriorly in the visual cortex, along the sequence of visual areas, the tuning of cortical neurons shifts from local, simple elements – such as lines, corners and contrast boundaries to more abstract and complex responses. The characteristic of neuronal tuning at the top of the visual hierarchy is still debated, but a growing body of work points to a strong holistic and configural element in the neuronal tuning. Thus, both human brain imaging and monkey single unit recordings reveal selectivity to various visual categories. Furthermore, direct demonstrations of holistic and configural properties have been consistently documented (e.g. Grill-Spector & Malach 2004). All these results converge in supporting the notion of a holistic, template – selectivity of neurons at the top of the cortical hierarchy.

The importance of these observations is that they offer a straightforward (although far from simple) feed-forward mechanism that could potentially solve the classic “binding” problem. In such a formulation, each neuron at the top of the hierarchy binds the local visual elements through converging feed-forward inputs from lower level visual areas. Thus, simple visual elements represented by the activity in primary visual neurons are ultimately “bound” into holistic templates corresponding to the unified perceptual images at the top of the visual hierarchy.

This is the classic sparse representation – popularly coined the “grand-mother” neuron hypothesis (Barlow 1972) – that keeps emerging under various modeling attempts.

5. A combinatorial explosion

While feed-forward hierarchical convergence could potentially solve the binding problem – it does not address the other fundamental problem raised earlier – the capacity of the visual system to generate trillions of distinguishable perceptual images. Note that even if we allocate a single “grand-mother” neuron to each perceptual image – there will not be enough neurons in the brain to account for the vast collection of all distinguishable conscious percepts. The solution, as indeed has been proposed by a number of distributed models, is that each neuron will be sensitive to a large number of different visual templates. We have previously proposed the metaphor of a “totem-pole” neuron to illustrate this notion (Gilaie-Dotan & Malach 2007). Basically the idea is that a neuron’s tuning can be envisioned as an *or*-function of many, likely millions, of different distinguishable holistic image templates – so that whenever we perceive a specific image, all the neurons that include this specific template in their tuning profile will fire.

The rules that organize the library of visual templates within each individual “totem-pole” neuron are not fully established. The discovery of face neurons as well as other object category neurons, suggests that the totem-pole libraries are organized according to a categorical principle. Furthermore, brain imaging and direct recordings from human cortex suggest that perceptual similarity may be another underlying principle – e.g. similarly looking exemplars within a category are more likely to be members of the same “totem-pole” library – i.e. co-activate a more similar set of neurons as compared to non-similar exemplars.

6. The ambiguity inherent in isolated responses

Regardless of the precise nature of the totem-pole tuning curves, the important principle for our consideration here is the fact that given the limited number of visual neurons – and the vast space of potential conscious percepts – each visual neuron must be tuned to a very large set of different visual percepts.

The problem with such broadly tuned neurons is that they are individually ambiguous – i.e. when an individual neuron is firing it is not possible to know which specific perceptual image this firing refers to. The solution to this

ambiguity, as indeed has been proposed by a number of models is in assembly representation. Thus, to take the totem-pole metaphor – if all neurons that contain a single common image, x , in their library are active together – this co-activation will unambiguously “point” to x as the perceived image. In contrast, those images which some of the neurons in the assembly are not sensitive to – will be ruled out.

Importantly, the library of each totem-pole neuron contains a subset of all the possible visual images, and furthermore, different totem-pole neurons have overlapping, yet different libraries. Such an arrangement generates a huge combinatorial space of possible images. Thus, if we consider the case of face representations, we have estimated indirectly that at least a million neurons are linked to the perception of each individual face (Levy, Hasson & Malach 2004). The number of different combinations that can be generated in assemblies consisting of a million neurons is truly colossal – and is more than sufficient to account for all possible perceptual images.

However, the principle of assembly coding of holistic “totem-poles”, which seems to solve both the problem of binding and that of the need for a huge storage of perceptual images, leads to the fundamental, often overlooked, problem of readout. As I have argued earlier, any model of conscious perception must consider the intrinsic perspective of the brain itself and not that of the experimenter collecting the data. The question that must be resolved then is that: what system within the brain itself, actually “knows” that a specific group of neurons – e.g. the assembly representing image x – has been activated?

7. Neuronal reflections

One can of course envision numerous mechanisms by which the activity of neuronal groups can be detected in the brain – however there is one unique process of information exchange that implements the self-readout concept in the most literal manner. I would like to put forward the conjecture that such process is the core dynamics underlying all conscious awareness states.

More specifically, I proposed that a visual percept emerges when the state of an activated assembly is made available to each of the member neurons that take part in that assembly. In other words, there is a special state in which each neuron in an assembly receives, directly or indirectly, a pattern of synaptic activations that reflect a *single and unique* activation state of all neurons in the assembly. Critically, the same single activation state is “felt” by all neurons simultaneously. This may immediately raise the worry of infinite regress – since the neurons that are affected by the state of the assembly are also the ones that constitute this state. However, as

has been demonstrated in many network models, such dynamics can settle, at least temporarily, into a stable consensus, in which all neurons are both affected and yet at the same time activate other members in the assembly. Note on the other hand, that such fully “democratic” distribution of information about a network state to all its members can be accomplished only in a recurrent network. By contrast, in feed-forward chains, the state of the system is not reflected in the synaptic inputs to all its member neurons.

To illustrate this conjecture in the context of visual perception – as was discussed above, the information inherent in the activity of isolated neurons is ambiguous and hence meaningless. However this ambiguity can be resolved by forming an integrated assembly. Here it is proposed that the means to achieve such critical integration is a unique dynamics in which the pattern of activity in the assembly is distributed to all its member neurons through reverberatory activity. Furthermore, when such recurrent dynamics reaches a momentary stable consensus – a perceptual image is created in the mind of the observer. We have recently proposed that such mutual activation among neighboring neurons can be metaphorically envisioned as a process of “ignition” (Fisch et al. 2009).

8. Experimental consequences

Although the ignition conjecture requires detailed modeling which is beyond the scope of the present chapter – even considering its more qualitative aspects can provide well-defined constraints and predictions. An important consequence of the hypothesis concerns the time it takes to distribute the information that an assembly has been activated to all its member neurons. Note that under the plausible assumption that neuronal signaling is encoded in inter-spike time intervals (instantaneous firing rates) – the speed by which neurons can transmit information about the activation of an assembly is critically dependent on the firing rate. Thus, a system in which the transition from inactive to active state consists of a change in firing from 1 to 5 Hz – will take a minimum of 200 msec for one signal iteration – a very long time when considering the fast response time typical of perceptual processes (Hung, Kreiman, Poggio & DiCarlo 2005; Thorpe 2005). Thus, high firing rate is critical for such assembly coding if it needs to comply with realistic speeds of perceptual responses.

A second consequence relates to connectivity – if information among neurons must be distributed rapidly – there should be dense anatomical connectivity between them to allow such rapid distribution. Furthermore, the critical information exchange should occur among those neurons that must

share information – e.g. have common “libraries”. Such neurons tend to cluster together anatomically – so that the connectivity must be mainly local, among close anatomical neighbors.

A third important consequence concerns stability – note that the content of the perceptual state is equivalent to the specific pattern of active neurons. To achieve a non-ambiguous percept, this pattern must remain stable for the duration it takes the neurons in the assembly to reach a consensus. This means that high firing rates must be sustained, and that a pair of spikes is not sufficient to signal the appearance of an assembly.

These are theoretical consequences of the ignition conjecture. The critical question is to what extent experimental findings confirm to these consequences. Below I will try to show that the ignition conjecture can account for a number of seemingly unrelated experimental observations from human and animal research.

9. What is so unique about the cerebral cortex?

The first experimental fact that needs to be taken into consideration is that the part of the brain which appears most closely associated with subjective awareness is the cerebral cortex. Thus both cortical lesions and electrical stimulation, and more recently TMS, have provided ample evidence to the critical link between this special organ and conscious awareness. But what makes the cerebral cortex so unique in its association with awareness phenomena? Below I will review the three most prominent aspects of cortical functional organization that emerge from recent neuroscience data.

Connectivity: Cortical neurons have dense local interconnections

A potential clue into the necessary conditions for awareness may be highlighted by examining what neuroanatomical features distinguish the cerebral cortex from other brain structures. While the feature that is classically emphasized is the laminated structure of the cortex, more recent neuroanatomical studies have also emphasized the dense recurrent connectivity (Douglas & Martin 2007). A powerful substrate for intense reverberatory activity, which has not received appropriate attention, is the dense “halo” of local connections which is particularly prominent in high order association cortex (See Figure 2, and (Amir, Harel & Malach 1993)) and contains numerous functional synapses. Such halos provide a dense substrate for reverberatory interactions at local scales – i.e. among neighboring neurons.

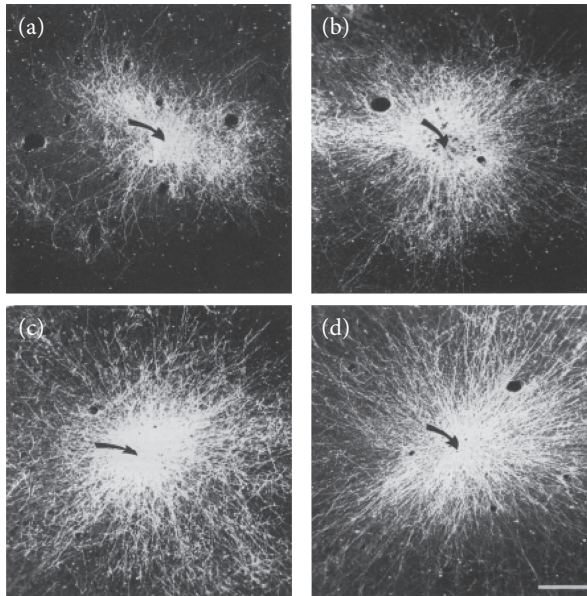


Figure 2. Top views of dense “halos” of lateral connections in primate visual cortex. Lateral connections can be seen emanating from neurons at the center of the tracer injection sites (arrows). Note the dense connectivity which increases in size from area V1 (a) through V2 (b) V4 (c) to 7a (d). The density of reciprocal interconnections can be appreciated when considering that such connectivity streams from every cortical site. Bar = 100 μ m. (Taken from Amir et al. 1993)

The crucial question is whether there is any evidence that ties such lateral connections to conscious awareness? Perhaps the most dramatic demonstration of the critical role of intrinsic cortical connectivity comes from split brain patients. In these individuals, the surgical interruption of connections between neurons located in the two hemispheres literally splits (or doubles) the subjective awareness into two separate, autonomous entities. This truly amazing manipulation of the awareness state has been documented in the classic studies of Gazzaniga and others, and provides a strong indication that the unity of consciousness depends on the availability of intra-cortical neuroanatomical connections. Furthermore, although it may seem that cross hemispheric connections represent long-distance neuronal interactions – this is not the case. Both their laminar distribution as well as their anatomical locations (e.g. the vertical meridian representation in retinotopic cortex) indicate that the inter-hemispheric callosal connections can be largely viewed as an extension of the local, intrinsic connectivity, that has been “stretched out” as a consequence of the of dividing cortical areas into two anatomically separate hemispheres.

High Activity: Subjective awareness is associated with high neuronal firing rates.

Recent single unit and brain imaging studies have revealed a close link between perceptual awareness and intense neuronal activity in high order occipito-temporal cortical areas. Most direct are those studies involving bi-stable stimuli, in which the physical stimulus is held constant while subjective percepts alternate between two states. Thus, In the Rubin vase-face illusion (Hasson, Hendler, Ben Bashat & Malach 2001) a clear correlation has been demonstrated between the emergence of face percepts and the level of BOLD activity in occipito-temporal cortex. Both single units (Sheinberg, Leopold & Logothetis 1995) and brain imaging (Tong, Meng & Blake 2006) have revealed a link between high neuronal activity and perceptual state during binocular rivalry. Finally, using a backward masking paradigm, brain imaging (Grill-Spector, Kushnir, Hendler & Malach 2000) and intra-cranial ECoG recordings (Fisch et al. 2009) reveal that high neuronal activity is tightly linked to crossing the consciousness threshold. These are just a few examples out of a long list of studies that lead to the conclusion that at least in high order visual cortex high levels of neuronal activity are tightly linked to the emergence of perceptual awareness.

Time: Subjective awareness requires a minimal duration (100–300 msec) of neuronal activity.

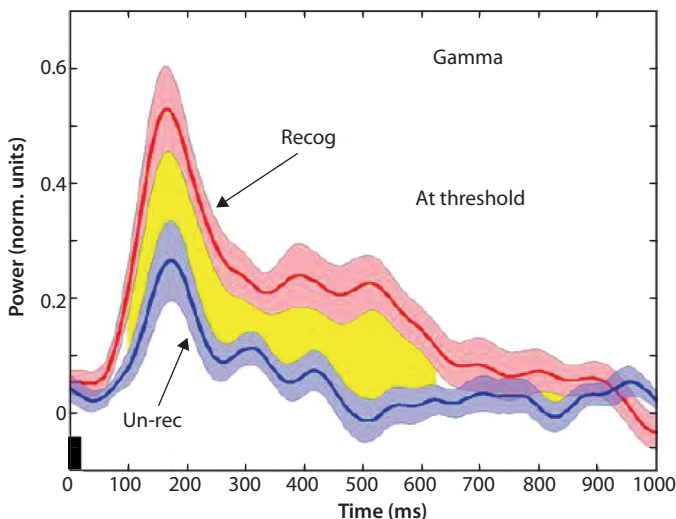


Figure 3. ECoG recording in high order visual cortex during visible and non visible trials in a backward masking experiment. Note the high and sustained neuronal activity during the visible (recog) state compared to the short and low amplitude response during the non-visible (Un-rec) trials. (Modified from Fisch et al. 2009)

Several lines of evidence point to the fact that subjective awareness can emerge only when neurons are active for a sufficient duration of time. Evidence for a minimal duration of neuronal activity comes from early studies of Libet, which demonstrated that, in patients undergoing surface stimulation of their primary somato-sensory cortex, sensory awareness emerged only after a rather prolonged (500 msec) period of cortical activation.

These results are also compatible with numerous visual masking experiments, in which the image is presented briefly, followed by a masking stimulus. These studies demonstrate that unlike non-visible stimuli, which lead to short lived neuronal activity, the neuronal responses associated with subjective awareness are typically sustained for 200–300 msec (Fisch et al. 2009). These results point to a minimal time that the neuronal signals need to “percolate” in the system to allow the perceptual event to be generated.

To summarize, converging experimental data point to high activity, dense connectivity and relatively long durations as critical for the emergence of perceptual awareness. These are precisely the same constraints that have been implicated as also critical for the formation of a consensus in neuronal assemblies proposed earlier. However, a fourth constraint – which is the most counter-intuitive and controversial concerns the locality of such assemblies.

10. Is consciousness a local or global phenomenon?

The crux of the assembly ignition hypothesis is that it allows for the possibility that subjective awareness could emerge locally – even within a small cluster of interacting neurons. Thus, as far as subjective awareness proper is concerned, the hypothesis does not require that the consensus about the assembly state will include large scale networks outside of the active assembly proper. The notion of a local assembly consensus as sufficient for perceptual awareness is of course quite counter intuitive – and in its purest form, can not in fact be verified, since the only means for the external world to be informed about the emergence of subjective awareness is through some kind of motor output – which by necessity requires that information, e.g. from the visual cortex, will eventually be transferred to some kind of behavioral output mechanism (Malach 2007).

The question, however, can be presented in a less extreme form: thus, we can consider whether the hypothesized cortical “ignition” underlying subjective awareness is a local or global phenomenon (Dehaene & Changeux 2011). To go back to the example of visual awareness – the question can be formulated more concretely as to whether the emergence of a conscious visual image entails an ignited assembly that includes not only visual neurons proper, but also high order,

e.g. frontal neurons that are associated with decision making, high order meta-cognition, and self-related aspects (Baars, Ramsoy & Laureys 2003).

It is important to emphasize that non-visual cortical areas as well as subcortical networks may also be involved in subjective visual experience. In fact non visual influences – such as efferent copy and attentional signals, arousal and even control of the sleep-wake cycle, are all critically important as modulatory and enabling factors – setting the excitatory level of the local neuronal assembly sufficiently high so that an ignition can occur. However, given the appropriate enabling factors – the conjecture is that the ignition-dynamics necessary for reaching a consensus state, i.e. high firing rate, sustained activity and reverberatory distribution of information can emerge completely locally in the visual assemblies proper. Unfortunately, the experimental data is still not conclusive on this point – and given its importance to the entire conjecture made here – this is where the local ignition hypothesis will stand or fall. However, although conclusive evidence is lacking – some experimental results that narrow the possibilities are already available.

The most straightforward evidence comes from the careful examination of neuropsychological cases. Thus, visual awareness appears to remain following extensive lesions of pre-frontal cortex, as well as disconnection of frontal cortex through lobotomy. In pre-frontal cortex, such lesions typically affect personality traits, decision making, self initiated acts etc. but subjects, even in cases of massive ablation or disconnection of frontal lobes appear to maintain their ability to consciously see visual images. This is in striking contrast with lesions in visual areas that lead to various and often dramatic forms of loss of visual awareness.

With regards to the more specific issue of the involvement of self-related networks in visual perception – this question has been examined in brain imaging experiments. Importantly, the results of such experiments reveal that pre-frontal “self” related cortical areas are actually *inhibited* during intense visual or auditory tasks. Activation of these areas was only apparent when subjects engaged in more reflective, introspective tasks (See Figure 4 and Goldberg, Harel & Malach 2006). These results indicate that during intense sensory experience, self-related, pre-frontal activity actually shuts off. This is probably due to the shift in the focus of attention towards assemblies related to the generation of perceptual images of the external environment, at the expense of more internally-oriented and self-related phenomenal states.

Other important structures implicated in subjective awareness are the parietal lobes. Indeed, damage to the parietal lobe in one hemisphere leads to sensory “neglect” which is undoubtedly one of the most striking cases of a selective loss of visual awareness. However, it should be noted that this dramatic effect is only apparent when the lesion is unilateral, i.e. affects one hemisphere. A bilateral lesion of parietal cortex paradoxically leads to a very different kind of deficit – the Balint

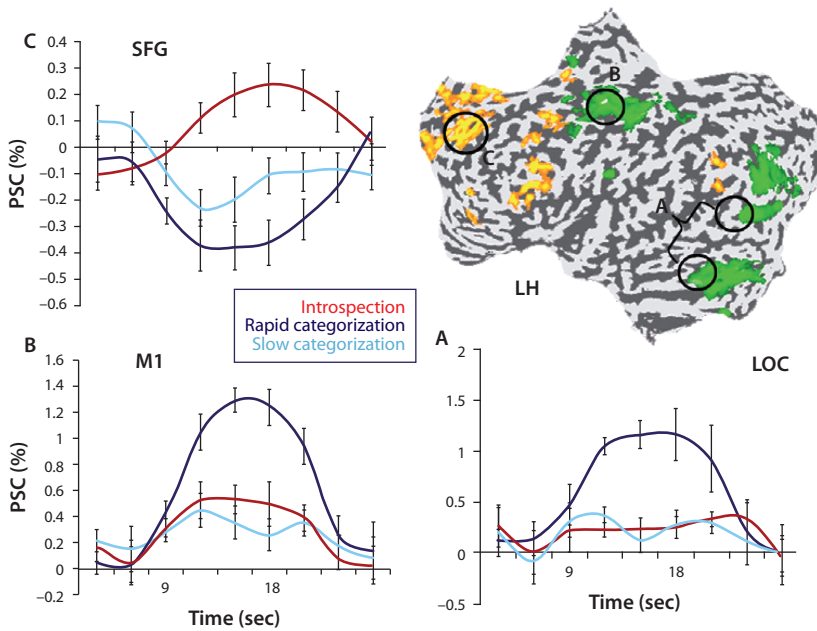


Figure 4. “Self-less” perception. Inhibition of self-related pre-frontal cortex during sensory perception. An unfolded map of the left cortical hemisphere. There is a strict segregation between self-related networks activated during introspection (orange colors) and sensory-motor networks activated during intense perceptual categorization (green). Note that activity in the superior frontal gyrus (SFG) was actually inhibited during intense perceptual categorization (SFG-inset) (Modified from Goldberg et al. 2006)

syndrome. In this syndrome, subjective awareness is largely maintained; rather, the Balint patient has difficulty in certain aspects of attentional control – such as shifting attention from one object or attending to more than one object at a time.

11. Is visual awareness dependent on higher levels of the cortical hierarchy?

When viewed in the context of the hierarchy of visual processing, it is assumed that in the visual domain this hierarchy leads to medial-temporal lobe structures such as entorhinal cortex and hippocampus. It thus could be argued that visual awareness depends on “readout” by these yet higher order structures. Indeed, highly selective and intriguingly invariant visual responses have been recently observed in single unit recordings in such structures in human patients. Furthermore, these responses were shown to be tightly linked to subjective awareness in backward

masking paradigms (Quiroga, Mukamel, Isham, Malach & Fried 2008). However, two aspects of the responses in these regions argue against their involvement in visual awareness. First – the responses begin at long latency – about 100 msec after the ignition of neuronal responses in visual cortex – (i.e. ~300 msec after stimulus onset) which, given the rapidity of visual perception is likely too long. Second, the firing rate increases typical of such medial-temporal lobe responses are rather slow – 5–10 Hz, again arguing against a sufficiently fast information exchange that is necessitated by ignition processes (Gelbard-Sagiv, Mukamel, Harel, Malach & Fried 2008; Quiroga et al. 2008).

To conclude, the evidence with regards to the locality vs. globality of assembly ignitions is still not conclusive, although the experimental results rather constrain the manifestations of such a global spread. At this stage, this question should be viewed as a prediction of the local ignition hypothesis. Thus, if it turns out that global ignition dynamics, as conjectured above, is an essential element in any sensory percept – this would constitute a refutation of the proposed conjecture.

The proposed model is of course very qualitative and sketchy, leaving numerous open questions. Thus, it is still not clear what is the biophysical manifestation of the signals that are integrated during ignition. Are there “echoes” of the ignition that do spread out to the rest of the brain and put the ignitions in their unique context?

The ignition model emphasizes the involvement of reciprocally active neurons in the perceptual dynamics. However, it leaves open the potential role that non-active neurons may play in the generation of a visual percept. A thoughtful and quantitative model that strongly argues for a critical role for such silent neurons has been put forward by Tononi in his integrated information theory (Tononi 2005).

Another particularly thorny difficulty for the local ignition model is that it does not offer an explanation for the apparent experiential unity of subjective awareness – since, in principle at least, multiple, separate ignitions may occur simultaneously. While one could envision a global brain mechanism that prevents such multiple ignitions, the evidence for or against such control is lacking. Thus, all these questions and many more remain open. Consequently, the conjecture put forward in this chapter should be viewed more as a working framework for future research than a conclusively proven model.

References

- Amir, Y., Harel, M. & Malach, R. (1993). Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual-cortex. *Journal of Comparative Neurology*, 334(1), 19–46.

- Baars, B.J., Ramsoy, T.Z. & Laureys, S. (2003). Brain, conscious experience and the observing self. *Trends in Neurosciences*, 26(12), 671–675.
- Barlow, H.B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1(4), 371–394.
- Dehaene, S. & Changeux, J.P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227. doi:10.1016/j.neuron.2011.03.018.
- Douglas, R.J. & Martin, K.A.C. (2007). Recurrent neuronal circuits in the neocortex. *Current Biology*, 17(13), R496–R500.
- Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., ... Malach, R. (2009). Neural “Ignition”: Enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron*, 64(4), 562–574. doi:10.1016/j.neuron.2009.11.001.
- Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R. & Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, 322(5898), 96–101.
- Gilaie-Dotan, S. & Malach, R. (2007). Sub-exemplar shape tuning in human face-related areas. *Cerebral Cortex*, 17(2), 325–338.
- Goldberg, II, Harel, M. & Malach, R. (2006). When the brain loses its self: Prefrontal inactivation during sensorimotor processing. *Neuron*, 50(2), 329–339.
- Grill-Spector, K., Kushnir, T., Hendler, T. & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, 3(8), 837–843.
- Grill-Spector, K. & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27, 649–677.
- Hasson, U., Hendler, T., Ben Bashat, D. & Malach, R. (2001). Vase or face? A neural correlate of shape-selective grouping processes in the human brain. *Journal of Cognitive Neuroscience*, 13(6), 744–753.
- Hung, C.P., Kreiman, G., Poggio, T. & DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749), 863–866.
- Levy, I., Hasson, U. & Malach, R. (2004). One picture is worth at least a million neurons. *Current Biology*, 14(11), 996–1001.
- Malach, R. (2007). The Measurement problem in consciousness research. *Behavioral and Brain Sciences*, 30, 516–517.
- Quiroga, R.Q., Mukamel, R., Isham, E.A., Malach, R. & Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 105(9), 3599–3604.
- Sheinberg, D.L., Leopold, D.A. & Logothetis, N.K. (1995). Controlling Perceptual Alternations During Binocular-Rivalry in Humans and Monkeys. *Investigative Ophthalmology & Visual Science*, 36(4), S668–S668.
- Thorpe, S.J. (2005). Ultra-rapid scene processing – Just how fast? *Journal Of Cognitive Neuroscience*, 10–10.
- Tong, F., Meng, M. & Blake, R. (2006). Neural bases of binocular rivalry. *Trends In Cognitive Sciences*, 10(11), 502–511.
- Tononi, G. (2005). Consciousness information integration, and the brain. *Boundaries Of Consciousness: Neurobiology And Neuropathology*, 150, 109–126.

From probabilities to percepts

A subcortical “global best estimate buffer” as locus of phenomenal experience

Bjorn Merker

A brain charged with guiding its body through a complex and lively world from a position of solitary confinement inside its opaque skull faces a set of functional challenges beset with inverse and ill-posed problems at every turn. Uncertainty and ambiguity therefore encumber all cortical labors, making probability distributions the natural medium of its disambiguating inferential operations. This chapter proposes that those operations take place unconsciously, in keeping with Helmholtz’ original suggestion, and that the functional logic of an inherently probabilistic cortex implies a need for an extracortical “global best estimate buffer” as a means to complete cortical sensory disambiguation through a definitive but ephemeral estimate of current sensory circumstances. It further proposes that the contents of that extracortical buffer are conscious, not by virtue of anything being “added” to buffer operations in order to “make them conscious”, but by virtue of the format of buffer contents alone: its dynamics issue in a nested arrangement placing an ego-center in perspectival relation to a neural model of body-world interactions. Finally, the organization of the higher order nuclei of the dorsal thalamus are scrutinized for their suitability to implement the putative global best estimate buffer, with particular attention to the possibility that the caudal reaches of the dorsal pulvinar might host its specifically sensory aspects, i.e. sensory awareness.

Keywords: ambiguity; architecture of consciousness; best estimate buffer; constraint satisfaction; format of sensory awareness; phenomenal content; probabilistic operations; pulvinar

1. Introduction

In an earlier publication I proposed that consciousness pays its way in the functional economy of the brain by unlocking the savings hidden in the mutual dependencies among target selection, action selection and motivational ranking

through multi-objective constraint satisfaction among them (Merker 2007, p. 70). Since the full scope of these savings is available only at a global interface between the three task categories, this would place consciousness at a late stage in the brain's operations, and suggests a subcortical implementation of key mechanisms of consciousness in sites of global convergence in midbrain and diencephalon. The claim was not that these key subcortical mechanisms suffice to implement the full phenomenal contents of adult human consciousness, but only that their organization is such as to plausibly sustain some forms of conscious content on their own, even in the absence of the cerebral cortex. Issues such as how and where our phenomenal experience of a fully articulated three-dimensional world populated by interpreted perceptual objects might be implemented, and whether any conscious contents are in fact implemented at the cortical level itself, were left open.

No doubt our cortical machinery is the source of much of our conscious contents. That does not mean, however, that the cortex must also be the site where those contents become conscious. They might well reach consciousness only by engaging a distinct system dedicated to conscious phenomenology, along the lines of what Daniel Schacter – without committing himself to its manner or locus of implementation – called a “conscious awareness system” (Schacter 1989, 1990). Such a system might in principle be extracortical, and receive its cortical information by direct, monosynaptic, projection from cortical pyramidal cells.

There are indications in information-theoretic analyses of specifically cortical modes of functional organization that suggest that an extracortical conscious awareness system may not only be eminently compatible with the way the cortex works but, as we shall see in what follows, it may well be a necessary consequence of its probabilistic mode of operation. The midbrain and diencephalic convergence sites already alluded to might have a natural role to play in such an arrangement, which as a whole would allow an unconscious cortex to contribute its information to a subcortical conscious awareness system on a moment to moment basis, and thus add its capacious temporal horizons to our being in time.

2. The ubiquity of sensory ambiguity and its absence from sensory consciousness

Enclosed in an opaque skull, the brain knows nothing of the world that surrounds it apart from irregular spike trains delivered at the proximal ends of its sensory nerve fibers. Its situation in this regard is unlike that of the neurophysiologist studying its operations. The neurophysiologist is able to monitor both neural responses inside an animal's brain and the real world stimuli that drive those responses. The

brain itself has no independent access to the stimuli activating its sensory nerves, and accordingly cannot correlate the two for interpretive purposes. Were it not for the fact that sensory fibres in the modalities of sight, hearing, and touch form spatially coherent bundles such that neighboring surface receptors are connected to neighboring patches of the brain's sensory maps, it seems unlikely that the brain would ever arrive at anything deserving the name of 'world' at all.

Yet even then, not only are the spike trains that arrive via these fiber bundles inherently noisy, they originate in receptor arrays confined by biological necessity to the surface of the brain's body, a body which – to complicate things further – is mobile with respect to the world those receptors respond to. The simple fact that receptors are confined to the body surface, i.e. are located at a remove from the sources in the world they are designed to monitor, encumbers the brain's task of interpreting their spike barrages with inverse and ill posed problems at every turn (Howe et al. 2006; Leopold et al. 2005; Marroquin et al. 1987; Pizlo 2001; Poggio & Koch 1985). An elementary example of such a problem is illustrated in Figure 1.

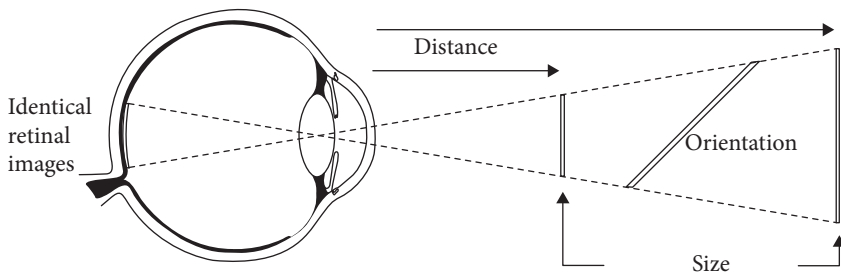


Figure 1. The geometry that relates retinal images to their sources in the physical world is projective, from three spatial dimensions to two. This eliminates one-to-one correspondence between images and their causes, saddling the brain's task of interpreting its sensory input with inherent ambiguity. Here a sample of an infinity of objects differing in size, distance or orientation that generate one and the same retinal image is illustrated. After a figure in Howe & Purves (2005)

Inverse problems abound in all stimulus dimensions: Is this a highly reflective surface under low illumination or a poorly reflective one brightly lit? Is this acute angle actually an obtuse one viewed sideways? A full list of inherent sensory ambiguities has never been compiled, but in the modality of vision alone it includes (but is not limited to) the relationship between our conscious percepts and stimulus dimensions such as size, distance, orientation, contour, texture, shading, illumination, reflectance, transmittance, motion, binocular disparity, solid shape, object groupings, and scene segmentation. These and analogous ambiguities in

other modalities are typical of the kind of challenges that routinely confront the brain's sensory operations.

Ambiguity is in fact the rule rather than the exception. So called ambiguous figures are merely stimulus situations that travel all the way to the end of the brain's ambiguity-resolving efforts without resolving, and thus reach consciousness to be *experienced* as ambiguous. All the rest, however, invariably start out encumbered by intrinsic ambiguity, but resolve in the course of their passage to consciousness through the brain's sensory systems. The fact that the ambiguities span a wide range of interpretive complexity means that they cannot be resolved at one level of analysis, and so impose a hierarchical organization on sensory systems. It is through their sophisticated labors that ambiguity eventually is resolved, and we are presented with a coherent sensory world composed of objects of definite sizes, distances, shapes, countours, colors, and trajectories. So thorough is this disambiguation that we normally are oblivious to the very existence of ambiguity, unless, of course, we happen to grapple with the problems of vision intellectually. Even then we penetrate the apparent unambiguousness only piece-meal and in the abstract, as it were.

Note the contrast already emerging between what is present to our sensory awareness and what the cortical sensory systems must labor with in order to supply us with unambiguous percepts. Their feats of thorough disambiguation in the face of the ill-posed nature of their sensory tasks is a crowning achievement of telencephalic evolution. It is yet to be rivalled by even our most sophisticated artificial systems, employing components that work at speeds some seven to eight orders of magnitude faster than biological neurons. Major efforts in the study of perception and computer vision have been devoted to unravelling the means by which our perceptual systems nevertheless accomplish their tasks in the face of inherently noisy and ambiguous input. Typically they show the utility and even necessity of applying prior constraints, regularization, and inference operations to the resolution of sensory ambiguity (Freeman 1994; Girosi et al. 1995; Knill & Richards 1996; Rao & Ballard 1997; Witkin 1981). And behind these devices there appears to lie one further, fundamental, key to the successful operation of the cortical sensory hierarchies: their adoption of a policy of honesty. Ambiguity and uncertainty is the reality they face, and they appear to represent it as such, by means of probability density distributions over alternatives mapped onto populations of neurons, starting with the cortical column itself (Anderson & Van Essen 1994; Deneve et al. 1999; Földiák 1993).

Once sensory information is framed in a format of probabilities and their associated uncertainties, it becomes possible to adopt a strategy of waiting for additional information to clear things up, because a probabilistic format renders information from various sources combinable within the limitations of neural

operations, and sometimes with near-optimal efficiency (see, e.g. Deneve 2008a, b; Ma et al. 2006). The “additional information” might come in the form of additional cues – say from a head movement or a second eye in the situation depicted in Figure 1 – or from prior constraints on the space of solutions, built into the interpretive system through evolutionary design or through experience.¹ In the latter case, past experience establishes an empirical probability distribution across the range of alternatives, which then acts to constrain interpretations on the next encounter with similar circumstances (see Edelman et al. 2002; Pouget et al. 2003; Purves et al. 2001). Probabilities are no less in evidence when additional cues are recruited to the task of disambiguation, since both human observers (for example: Knill & Saunders 2003) and macaques (Gu et al. 2008) take stimulus uncertainty into account when cues are being combined (see Ma 2010 for a detailed treatment).

This all assumes, of course, that there is a capacious neural space available to probabilistically map the range of alternatives across populations of neurons for the vast range of sensory items and events our sensory systems are manifestly capable of resolving. This would seem to be a principal task of the mammalian cerebral cortex, and one of the selection pressures behind its evolutionary expansion. By extracting sensory probability distributions from its constant afferent bombardment, it can then bring these “acquired priors” to bear on ambiguous aspects of afferent activity at any given moment (Anderson & Van Essen 1994; Földiák 1993; Hinton & Brown 2000; Knill & Richards 1996; Kording & Wolpert 2004; Ma et al. 2006; Pouget et al. 2003; Zemel et al. 1998; see also Hinton & Sejnowski 1983, which foreshadows many a subsequent development in this area). Indeed, probability density distributions may furnish the common currency of cortical operations (Andersen & Van Essen 1994), to be collapsed onto estimates only at decision time (Ma et al. 2006).

That is, to maintain ready combinability of the contents of its sensory operations, the cortex may exercise a “wisdom of suspended commitments” implemented in probabilistic terms throughout its sensory hierarchies. It would be, as it were, the neural version of the programmer’s or modeler’s “principle of least commitment” (Marr 1976). The strategy of postponing the derivation of sensory estimates to a late stage of operations is additionally supported by the fast propagation of firing rates that is available to the cortex by postponing full temporal integration

1. To a first approximation, constraint by innate design works by not representing the full range of ambiguity – providing, as it were, instant certainty at the price of permanent naivete (think of the frog moved to snap at small moving objects irrespective of their edibility) – while an experience-based constraint solution must provide a representational space covering the full range of ambiguity as a means to extract from experience a probability distribution of alternatives across that range.

to the last and final stage of a propagation path involving multiple synaptic relays (see van Rossum et al. 2002 for details). As van Rossum and colleagues show, the lack of precise firing estimates in this fast propagation mode is no bar to systems operations. The two “postponements” (of sensory estimates and firing estimates) may be functionally linked, and suggest a hitherto unappreciated aspect of cortical operations, namely that they may perform sophisticated, capacity-intensive, but preliminary operations for an ultimate, potentially extracortical, integrator of cortical labors.

Our sensory consciousness clearly bears the mark of taking place at the postponed (read-out, estimate) stage rather than at the postponing, cortical, stage of the brain’s operations. Under ordinary, non-laboratory-contrived, circumstances we are surrounded by definite objects rather than ambiguously or probabilistically defined ones. Objects are distinct and well defined on all the seven basic dimensions of vision (Adelson & Bergen 1991), and together they surround us in the form of the well articulated and interpreted three-dimensional extended panorama we call the world. This commitment of sensory *consciousness* to exclude ambiguity and tentativeness from our conscious percepts appears to belong among its more basic characteristics.

Indeed, sensory consciousness appears constitutionally incapable of truly representing stimulus ambiguity. When viewing conditions are restricted to remove cues to size and distance – say viewing a luminous phosphorescent disc monocularly with light-adapted eye in darkness – we still see an object of definite size floating before us. That impression amounts to a bluff: no veridical cues underwrite it, yet we are presented with a definite percept nevertheless. When a figure contrived to contain unresolvable ambiguity fails to resolve – say a Necker cube or Rubin’s faces-vase figure – it is not admitted to consciousness as it actually stands, namely as a single shape representing two or more equally valid objects concurrently. Rather, it is admitted only after being dressed up in the sensory fiction of a leisurely alternation between them, one sensory interpretation and appearance at a time, an experience for which there is no warrant in the sensory afference (which remains constant throughout). In this desperate stratagem the system is only adhering to, and thereby revealing, its normal mode of operation, which is to present the inherently ambiguous and uncertain to consciousness as something singular and definitive.

The categorical refusal of sensory consciousness to admit ambiguity suggests the operation of an iron-clad constraint of some kind, or even a capacity limit. The question is what might account for it. For one thing it is hardly conceivable that the trivial amount of pattern content that defines a Necker cube or Rubin’s faces-vase figure should tax cortical pattern-representing capacity to such an extent that its two aspects could not be accommodated within it. For another, we

know that information regarding the suppressed alternative can be recovered from a number of loci and levels in the cortical hierarchy (see, e.g. Alais & Parker 2006; Andrews & Blakemore 1999; Fang & He 2005; Jiang & He 2006; Sterzer et al. 2008), the clear cortical signs of stimulus dominance in such cases notwithstanding (Blake & Logothetis 2002; Leopold & Logothetis 1999). And finally, why should a cortex operating by means of probability density distributions balk at representing such figures to us in the form of balanced probabilities when that is the reality of the stimuli present to the senses?

There seems to be a sharp difference, then, between the cortical mode of dealing with sensory information and the contents of our sensory consciousness (cf. Velmans 1991; see also Kornmeier & Bach 2009). The former operates in probabilistic terms, deals in uncertainty throughout its operations, and conducts numerous distinctive operations on the sensory array simultaneously and in parallel. Our sensory consciousness, on the other hand, presents us with a well-defined, unitary sensory world free of ambiguity and maybes, even to the point of being incapable of representing genuine ambiguity. Cast in the medium of probability distributions our world would presumably look more like a mix of cloudy and drifting wisps bounded not by sharp surfaces but by the ever thinning gradients of statistical distributions rather than like the world that in fact surrounds us.

The contrast is compelling, but need mean no more than that the locus of our experience is not cortical, but extracortical, and works in the medium of estimates that must at some point be generated from cortical probability distributions if they are to fill purposes such as the control of behavior. On this view it would be estimates that reach conscious representation while the cortical realm of probability distributions, engaged in the inferential work of sensory disambiguation that feeds the estimates, would be unconscious. For numerous cogent examples of presumably cortical operations taking place without awareness, see Velmans (1991). Helmholtz (1867) called the inferential processes of sensory systems unconscious in good agreement with this contrast, to be further explored in the remainder of this chapter.

3. An orthogonal order of estimates and the ‘precedence of the gaze’

The crisply defined, multimodal, three-dimensional object world we perceive around us in the form of a unitary, panoramic and movement-stabilized arena for our actions owes both the sophistication and the sharpness with which its objects and events are defined to the flux of distributed probabilistic interactions that plays across the cortical sensory hierarchies. It is through this process that noise and ambiguity are cleared from an inherently ambiguous sensory afference, as if by a

filtering action. It does not follow, however, that the filter also must host the contents of our sensory consciousness to which it contributes its labors. Rather than to serve as an arena for settling ambiguities through wide-ranging informational interactions – as assumed in workspace or blackboard-type theories of consciousness – sensory consciousness appears to occupy a functional stage subsequent to the many interactive operations necessary to resolve sensory ambiguity and uncertainty. Its role at this late stage of operations may be predicated on a general difficulty encumbering the logistics of disambiguation, as follows:

The interpretation of a local sensory feature depends on its relationship to neighboring regions of the sensory space, and ultimately on the segmentation and interpretation of the scene as a whole. This places later stages of sensory operations in a superordinate position with respect to earlier ones, and puts a premium on interactive operations (see, for example, Lee et al. 2009; Stephen & Mirman 2010). Higher-order groupings may relegate a particular low-level item to insignificance or may even add local features where none in fact exist (e.g. Kanizsa triangle). Similar considerations hold for the recruitment of additional cues to the task of disambiguation: they may reside anywhere along the sensory hierarchy, and there is no knowing at what point in progress up a hierarchy a crucial cue may be encountered. It may even reside in another modality altogether (e.g. McGurk effect). The logic of sensory disambiguation is contextual (Bar 2004; Edelman 2002; Schwartz et al. 2007; van der Kooij & te Pas 2010), and contextual embedding is nested, i.e. broader, more general contexts encompass narrower, more specific ones (Merker 2004, p. 562).

For all these reasons it is advantageous for the cortex to postpone the precipitation of estimates as long as possible, because doing so allows additional cues and constraints to be recruited to the disambiguating process (meanwhile decision pressure mounts by the fractions of seconds, as we shall see). Ideally, no candidate interpretation at any sensory level should be regarded as definitive until “all the facts are in” (i.e. the analysis is complete; see Edelman 2002, p. 128, ‘meaning holism’). Only then is it safe from being overturned by additional contextual constraints or cues, and ready to find its place in a comprehensive scene interpretation. But how is any disambiguation at all possible under such circumstances, circumstances which would seem to make the resolution of any given ambiguity dependent on the resolution of all ambiguities?

There are two principal requirements for mastering this logistical dilemma. One is that information in all parts of the cortex and across sensory modalities must be capable of interacting rather swiftly and directly, in but a few steps of cortical inter-areal projection, so that cue combination and constraints can be recruited to the task without time consuming searches. Cortical connectivity helps fill this requirement through the bidirectional organization of cortical

sensory hierarchies. It makes the frontolimbic-hippocampal domain of the cortical inter-areal connective graph a “superhub” of a small-world connectivity that amounts to a neural countercurrent system (see Merker 2004 for details). To enable swift and wide-ranging interactions this connectivity must employ a shared data format, yielding the sub-requirement that, as suggested by Pouget and colleagues (Ma et al. 2006), probability density distributions not be collapsed to estimates within it.

The other principal requirement is imposed by the directional dependency of disambiguating operations within the cortical hierarchy, to the effect that higher levels take precedence over lower ones in the resolution of ambiguities. As we just saw, no candidate sensory alternative can be considered safe from being overturned until “all the facts are in.” This means that somehow a global scene interpretation is required for any disambiguation to be final at any level, i.e. definitive disambiguation can take place only in the form of a *global best estimate of the current scene as a whole*. Such a global best estimate clearly lies implicit in the totality of the probability distributions hosted by the cortical sensory systems, but to make it explicit would require that those probability distributions be reduced to estimate form. And, as just noted, since those distributions serve as the common currency of cortical interactions, the cortex would be ill advised to attempt to perform this reduction within its own operations.

We all know what the required global best estimate of the current sensory scene as a whole is like because – this is my claim – it is nothing other than the actual contents of our sensory consciousness, i.e. the world in which we find ourselves every second of our waking hours. Yet hoist with its own petard of a probabilistic data format, the cortex finds itself unable to precipitate this global best estimate within its own confines by its need to keep that data format patent for the wide-ranging cortical sensory interactions that take place by means of it. An elegant solution to this new dilemma would be to leave the cortical distributions as they are and to reduce them to a global estimate only extracortically, as suggested by Pouget and colleagues. They instantiate this suggestion in their formal treatment of the conversion of cortical probability distributions to estimates in the projection of parietal area LIP to the superior colliculus in the midbrain (Beck et al. 2008). In brief, a local attractor network in the intermediate layers of the superior colliculus, driven by parietal cortex area LIP, rapidly settles on a maximum-likelihood estimate of stimulus position for triggering a gaze displacement. Area LIP in turn is fed by extrastriate area MT, both of these employing probabilistic data formats.

This demonstration supplies proof of principle regarding the feasibility of dividing neural labor between a probabilistically operating cortical sensory hierarchy and a subcortical site where estimates are derived from cortical probabilities

for the control of behavior. The system modelled by Beck and colleagues is of course a special purpose one. Its scope is limited to one cortical area's contribution to the determination of stimulus position in the superior colliculus, and does not address the wider issue of deriving a global best estimate of the scene within which that stimulus is defined. Yet the particular behavior subjected to formal analysis in their treatment – gaze shifts – is a felicitous one, because gaze movements are more than simply a sample of behavior like any other.

Rather, movements of the gaze or orienting movements play a primary and superordinate role in the brain's orchestration of behavior. Whether appearing as an endogenously generated gaze shift integral to an ongoing task sequence (Hayhoe & Ballard 2005) or in the form of an orienting movement triggered by a salient sensory event in any modality (Sokolov 1963), the “very next act” is almost invariably *led* by a gaze shift (Johansson et al. 2001; see further below). The initial sign of evolving action in the logistics of the brain's control of behavior is typically a gaze movement. This is true even when the optic nerves are unconnected to the brain, and the eyes themselves therefore supply the brain with no information (Gamper 1926). We are concerned, in other words, with the earliest signs of the brain's phasic behavioral (motor) output, of which none is earlier or more rapid than movements of the gaze and other components of orienting behavior. Their mechanisms, in turn, are intimately related to those of spatial attention (Corbetta et al. 1998; Lovejoy & Krauzlis 2010; Rizzolatti et al. 1987).

Peripheral vision suffices for many purposes of ambient orientation and obstacle avoidance (Marigold 2008; Trevarthen 1968; Zettel et al. 2005) and normally proceeds without engaging focal attention. Yet when locomotion is initiated or redirected towards new targets or planned while traversing complex terrain, the gaze leads behavior by fixating strategic locations ahead (Marigold & Patla 2007). The strategy applies no less to reaching and manipulative activity, down to the split-second details of their trajectories. Fine-grained behavioral monitoring of eyes and manipulative activity in the laboratory has disclosed that the gaze *precedes* directed movements of arm and fingers by landing on behaviorally significant points *ahead* of the body part movements (Johansson et al. 2001). Arm and fingers follow the agile movements of the gaze *as if attached to it by elastic bands*. The coupling of arm or hand to the gaze appears to be the brain's default mode of operation (Gorbet & Sergio 2009), and so called gain fields (Andersen & Mountcastle 1983; Chang et al. 2009) can be likened to the “elastic bands” in the analogy just used.

The leading role of gaze shifts or orienting movements in the orchestration of behavior makes them the brain's primary and ubiquitous output. It is through them that the brain most immediately expresses its behavioral priority of the moment, and it is to them that it yokes the recruitment of the remainder of its

motoric equipment in implementing actions. This “precedence” or priority of the gaze means that its targeting movements must be informed by the brain’s best estimate of the sensory scene lest gaze shifts, and with them all the rest of behavior, be less appropriate and efficient than they need be. It is accordingly on the paths leading to the brain’s control of the gaze that we should expect to find an implementation of a global best estimate of the current sensory situation.

The gaze moves through coordinated movements of eyes and head. Empirically there is close covariation in the parameters of the respective movements of eyes and head during orienting gaze shifts, implying that they share a common neuronal driver (Guitton et al. 1990). The most caudally located premotor site with neuronal characteristics appropriate for exercising this unitary control of the gaze is the intermediate layers of the superior colliculus in the roof – tectum – of the midbrain (Freedman & Sparks 1997; Munoz et al. 1991). This puts a caudal bound on the targets of cortical projections among which to search for an implementation of the brain’s global best estimate of its sensory circumstances. If we want to explore candidate implementations of the brain’s global best estimate, it is therefore along the cortical approaches to this midbrain terminus that we must search. The superior colliculus, then, is the caudal-most terminus at or before which such an estimate must be realized.

4. Design features of a global best estimate buffer

Even a search space narrowed to the compass just suggested leaves a vast territory of elaborate subcortical structure where the presumptive best estimate might be implemented. It might be narrowed down further by trying to match candidate mechanisms against the attributes they must exhibit in order to qualify as a sensory awareness system. There is no dearth of demanding functional requirements to draw on, but before embarking on specifics, it might be useful to have a rough and ready idea of the minimum number of neurons required to represent – at full resolution – the momentary content of human sensory awareness in all its richness.

4.1 The neural numerology of sensory experience

It tends to be assumed that the neural resources required to implement consciousness generally (Tononi & Edelman 1998) or visual sensory consciousness more specifically (Pollen 2011) are so vast that the cerebral cortex is obligatory for the task. This ignores, however, the distinction drawn in the previous section between, on the one hand, the inferential operations the brain must engage in to

penetrate the thicket of inverse problems it is faced with and, on the other, the task of condensing the outcome of those operations to a definitive estimate, the result of which we have provisionally identified with the contents of our sensory consciousness. The former are vast and resource-intensive indeed, accounting for the elaborateness of the cortical machinery devoted to them, but they may take place unconsciously (Velmans 1991). By contrast, the task of deriving an estimate from these operations can make do with far more modest neural resources since the heavy-duty labor has already been performed in the cortical preliminaries, leaving only a reduction to estimate form for the conscious awareness system to accomodate. The question, then, is how compactly the momentary contents of sensory consciousness might be represented in a neural medium.

Vision is arguably the most demanding of the senses in this regard, being the dominant sense in terrestrial mammals and the general framework with which the other senses are aligned, not least because of the issues considered in the discussion of the precedence of the gaze. How many neurons are required to represent the momentary state of our perception of the world in as far as its purely visual sensory aspects are concerned? The question, in other words, concerns only the purely sensory content of our experience, in all the detail and nuance available in a momentary visual “snapshot”, but excluding for now what we know about what we see. Any answer will depend on numerous assumptions regarding how such an image might be represented in the medium of neurons and their activity, and we shall make the conservative assumption that they do so directly, by serving as individual pixels of an image space. How many such pixel neurons would be required to fully match our perceptual visual experience?

Well-founded estimates for vision as a whole have not been made, so we must extrapolate from estimates for restricted cases. One such calculation concluded that a monochromatic, monocular view, taking differential acuity across the visual field into account, requires some 164000 picture elements to match a human visual “snapshot” (Royer & Schwartz 1999; see also Watson 1987). Another investigator asked how many neurons would be required to represent all the spatial and chromatic detail we can see in an image, and answered “probably many fewer neurons than are contained in the lateral geniculate nucleus” (Lennie 1998, p. 900). So far these estimates are at least compatible, since the lateral geniculate contains some one million neurons.

That is also roughly the number of fibers composing the optic nerve and tract, which as sole conduit of visual information to the brain sets the limit on the brain’s running access to veridical visual information. The capacity of that conduit is likely to include some redundancy in support of cortical disambiguating operations, so a best visual estimate of the situation in the physical world that is driving those fibers may well be framed in a space of less capacity than the conduit itself.

For reasons we shall presently come to, the depth dimension of visual experience need not add heavily to the modest numbers so far considered, which means that somewhere between 10^5 and 10^6 neurons might suffice to render a realistic image of our visual surroundings.

Sensory experience as a whole is multimodal, so the visual estimate must be expanded to accommodate the remaining senses. They are, however, unlikely to add comparable numbers to the total estimate, because they share the spatial framework of visual perception, being “contained within it” as it were. This leaves a few million neurons as a rough and ready estimate for the total representational requirements of full moment-to-moment sensory awareness, once its preliminaries have been accomplished elsewhere. This crudely derived figure suffices to dispell the impression that vast representational resources are needed to accommodate sensory awareness. With its 16 billion neurons, the human cortex exceeds that number by between three and four orders of magnitude. One is thus free to explore whether less capacious – subcortical – structures are capable of filling the functional requirements of the brain’s sensory awareness system. We turn next to a consideration of such functional requirements.

4.2 Our panoramic, nested, ego-centric, three-dimensional visual world

Conscious percepts such as the sight of a chair or of storm clouds gathering on the horizon do not come to us piece-meal, but framed by a magnificently organized macro-structure, the format of our ordinary waking visual experience. The striking structural specificities of this macro-structure supply a ready-made set of “format requirements” to be met by a neural candidate for the brain’s global best estimate mechanism, and we shall draw on them for that purpose. They include its panoramic, three-dimensional and ego-centric (viepoint-based) geometry, its tripartite nesting of visual egocenter, visual body, and visual world within this geometry, the stability of the visual world despite eye, head and body movements that set the retina – sole mediator of our visual impressions of the world – jerking and flailing about in space, and finally the artifice of having the visual ego-center access the visual world from inside the visual body via an empty cyclopean aperture in the upper part of that body’s face region. We shall consider these aspects in turn.

Besides its panoramic extent in the x-y plane, our visual world has conspicuous depth; it is ineluctably three-dimensional. The depth dimension extends from the visual ego-center located inside the head (behind the bridge of the nose: Hering 1879/1942; Howard & Templeton 1966; Roelofs 1959) to the horizon and no further. Visually, we do not inhabit an unbounded space, as would be the case if our visual system performed a full inverse optics operation on its retinal input (Hatfield 2003). Such an operation would give us back the boundless structure of

physical space. Instead our visual space is a foreshortened version of the physical, geometrically shrunk into a finite ovoidal volume (cf Lehar 2002; Trehub 1991). It meets us directly in the shape of the dome of the sky ending at the horizon. No visual events are *perceived* as taking place beyond this concave boundary, though they may very well be *known* to do so. Those that in fact do take place beyond it are experienced as if taking place *on* it, most tangibly in the spectacle of the starry heavens on a clear dark night. The basic circumstances of this foreshortening are illustrated in Figure 2.

This bounded, three-dimensional image of a boundless world is partitioned into two major image compartments: a proximal visual body and a distal visual world

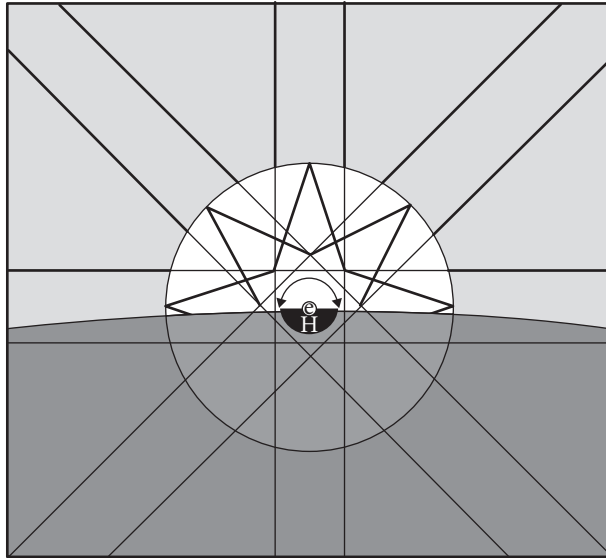


Figure 2. Highly schematized depiction of a bird's eye view of a simplified human observer, represented by a head (labeled H), standing at the intersection of four railroad tracks extending to infinity on a boundless plain. This observer has a limited field of view set by the eyes' bony orbits. Unseen parts of the scene are shaded. In the observer's conscious perception of this scene, the actual geometry of the landscape, in which parallel lines never meet, is remapped into a finite perceptual space bounded by the horizon (and the dome of the sky), at which parallel lines meet. It is represented by the large circle in the figure. To avoid cluttering the illustration, the *perceived* railroad tracks (depicted as heavy black lines inside the unshaded portion of the circle marking the observer's horizon) have been truncated at their close approach to the observer, whereas the perceived reality is that they "bend" around the observer's position, as it were (see Hatfield 2003). "e" marks the visual ego-center, and the double arrow marks the visual field of view, or "visual aperture." Image by Bjorn Merker licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

(the former is represented only by the head, marked H, in Figure 2). Within this partitioned space of experience all global displacements of the world's optical image on the retina are registered as movements of the body (as a whole or of its head and eyes) within an immobile and stationary world. The visual world compartment all the while exhibits a perfectly stable (non-translating) visual surround. What we see is obviously not based directly on retinal images, but on a derivative synthesis between retinal information and information about bodily movement, including those of the gaze. By segregating the synthesis into a body compartment nested within a world compartment, and letting the body compartment reflect all sensory aspects of self-motion in the form of movement of the body and its parts relative to a stationary world compartment, the latter is left as a stable arena for our actions in visual awareness.

Figure 3 illustrates in cartoon fashion how in such a scheme the compound sensory change occasioned by a gaze movement might be represented as the swivelling action of the visual head with its cyclopean visual aperture relative to a stable visual world. It would seem easier, in other words, to remap the coordinates of a central head relative to the surrounding world, than to remap the entire world relative to the head. Notice that this scheme implies that the neural space hosting the world-compartment represents a full 360 degree surround, of which the visual system in species with frontally placed eyes only samples a limited sector through its visual aperture at any given time. The auditory system, meanwhile, utilizes the full surround extent at all times and the space, of course, is multimodal.

The separation of body and world supplies such a scheme with an at least implicit dimension of depth. More generally, our sense of the world's three-dimensionality owes much to monocular cues such as perspective, occlusions (object interpositions), and gradients of texture and shading (Pizlo et al. 2008; Witkin 1981). The prepotency of monocular cues applies to the three-dimensional appearance of objects no less than to our visual surround as a whole (Li et al. 2009). Thus the world still looks three-dimensional through one eye only, and we maneuver and manipulate quite competently that way even in cluttered and complex environments. Binocular vision is obligatory only at the limits of resolution in depth, such as for threading a needle. Moreover, the views of the two eyes enter visual awareness in fused form, as if seen from a single, cyclopean vantage point (Julesz 1971). It is safe to assume, therefore, that costly operations like matching image points for stereo disparity lie among the cortical preliminaries to conscious perception, and need not burden the global best estimate buffer except in as far as it must have the capacity to register the *results* of multiple depth cues.

The synthetic global framework of nested sensory domains is held together by the ego-centric spatial frame of reference (coordinate system) for which the

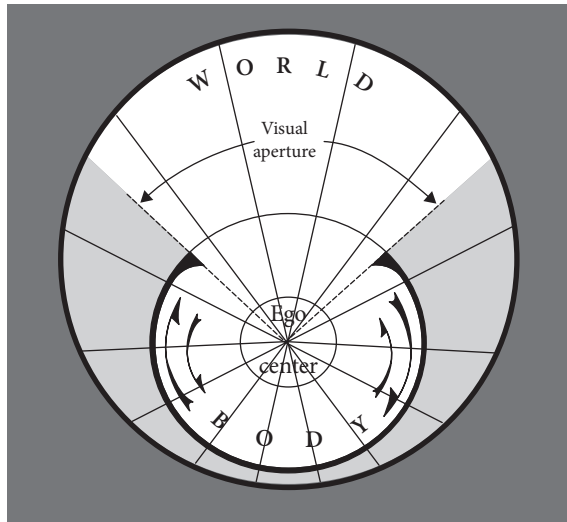


Figure 3. The principle of compensatory representation of self-motion in the egocentric space of visual awareness, depicted for a simplified gaze movement in cartoon fashion. For illustrative convenience, the gaze moves as a whole, without separate representation of eye movements. The spatiotopic and ego-centrally organized neural space is partitioned into three nested zones: egocenter, body map (here limited to a head representation) and distal world map. In this partitioning, global sensory motion contingent on active gaze displacement is represented as a swivelling action (curved single arrows) of the visual head within the full 360 degree implicit coordinate space. At the end of the movement a new sample of the distal visual world fills the visual aperture. The central placement of the ego-center lies closer to that of audition than to that of vision, a placement motivated by the fact that even in the more realistic case of independent eye movements the limits of the visual aperture are determined largely by the bony orbit of the eye, which (like the ear) is fixed to the head. Thus, a 45 degree rightward saccade of the eyes extends the visual field by far less than 45 degrees to the right. This makes head movements the major means for moving the visual aperture, and cephalomotor and vestibular signals central to the compensatory scheme (see Footnote 8). Given central magnification in vision, which foreshortens the magnitude of peripheral positional displacement on retinotopic maps during saccades, saccades are more a means for moving the “spotlight” of central, high-resolution vision than for moving the visual field as a whole. Image by Bjorn Merker licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

visual ego-center lodged inside the head of the visual body supplies an origin. That coordinate origin is located singly and centrally several centimeters behind the bridge of the nose, and it is from this point that we look out – in visual awareness – along straight lines of sight, upon the neurally modeled world that surrounds us, through what appears as an empty aperture in our

heads (Cox 1999; Hering 1879; Howard & Templeton 1966; Neelon et al. 2004; Roelofs 1959). It is the ego-centricity of this organization of sensory space that lends all visual objects their sidedness and handedness (for which see James 1890, vol. II, p. 150; Green 2003). It is impossible for us to *see* the objects of the world allo-centrally, we can only *know* about their relation one to the other mediately, via the irremediably perspectival view created by viewing the world from a point (Green 2003).

A visual ego-center that has free visual access to the visual world surrounding the body from a location inside that body's head is of course a sensory fiction of monumental proportions. This because we know very well that all paths from such a point behind the bridge of the nose in actual fact must pass through body tissues of various kinds if they are to reach the world around the body, yet visually there is no sign of these intervening tissues. Their place is taken instead by the neural fiction of an empty oval cyclopean aperture occupying the upper part of our face region. It is illustrated in a monocular view in the classic drawing by Ernst Mach, reproduced in Figure 4.



Figure 4. Ernst Mach's classic rendition of the view through his left eye. Inspection of the drawing discloses the dark fringe of his eyebrow beneath the shading in the upper part of the figure, the edge of his moustache at the bottom, and the silhouette of his nose at the right-hand edge of the drawing. These close-range details framing his view are available to our visual experience, particularly with one eye closed, though not as crisply defined as in Mach's drawing. In a full cyclopean view with both eyes open the scene is framed by an ovoid within which the nose typically disappears from view (see Harding 1961, for a detailed first person account). Apparently Mach was a smoker, as indicated by the cigarette extending forward beneath his nose. The original drawing appears as Figure 1 in Mach (1897, p. 16). It is in the public domain, and is reproduced here in a digitally retouched version, courtesy of Wikimedia (http://commons.wikimedia.org/wiki/File:Ernst_Mach_Innenperspektive.png)

These various design features of our sensory (visual) awareness are far from arbitrary. Rather, they summarize in a single geometric image-space the principal relations of the physical constellation of receptor arrays and their relative motions with respect to each other and the world that feed the cortical sensory hierarchies their primary information. Physically – though the retina would not know it – the world *is* stationary during an eye movement, yet the brain is fed a retinal sweep to be disambiguated as to its true source by the brain, and so on for combined eye and head movements, locomotor translations, etc. The incessant motion of receptor surfaces occasioned by movements of the body and its parts feed a constant stream of sensory variance to the cortex to be deciphered by its inferential machinery.

The ubiquity of gaze and body movements encumber the brain's sensory interpretation task with many of its central problems, such as sensory effects of self motion within modalities and spatial misalignment between modalities when the eyes move in the head, the head moves on its body and body parts move relative to one another. Yet at the same time facts such as the physical coupling of skeletal joints, the fact that eyes and head have centers of rotation and are physically attached to the body, and that body motion is a more common source of sensory displacement of the world than, say, earthquakes, embed correlations and clustering of correlations into the sensory variance caused by body movement. Cortical information is riddled, in other words, with movement-contingent clusters of correlated variances, and these, as everything else the cortex is fed, serve as grist for its mill of inferential disambiguation and memory-based accumulation of acquired priors. At their most global these correlated variances, generated as they are by a receptor-equipped physical body moving through a physical world, amount to a grouping of cortical information into the sensory super-clusters of "body" and "world" (see Philipona et al. 2003, 2004).

As the most general sensory invariant hidden in cortical probability distributions this body-world distinction, along with subsidiary invariants generated by physical constraints on the articulation of body parts, must figure at numerous points in the cortical operations that issue in its global best estimate of its sensory circumstances. In fact, the details of movement-contingent sensory change supply the brain with an accurate source of information for the reconstruction of its receptor arrays' trajectories through the world, and through them of the body's movements relative to that world. The *best* estimate of the causes behind the sensory complexities faced by the cortex trying to make sense of an afference generated by sensors mounted on a mobile and multiply jointed body is bound to be a neural model or simulation of the outlines of the very physical constellation that generated those complexities in the first place, namely a multijointed and mobile body

equipped with eyes, ears, etc. interacting with a body-independent world. And that, of course, is the way in which our situation is presented to us in sensory awareness, giving us our body and our world as neural contents of the best estimate buffer.

From this point of view it is even possible to understand the necessity of the neural fiction of a cyclopean aperture. In the egocentrically organized space through which body and world are related in the neural model, the body surface must be included as an object in the visual mapping of the world. Such inclusion is unproblematic for distal parts of the body. They can be mapped into the best estimate neural model in the same way as any other veridically opaque visual objects in the world. However, in the vicinity of the egocenter itself, persistence in the veridical representation of the body as visually opaque would block egocenter visual access to the world, given the egocenter's location inside the head representation of the body (where it must lodge in order to exploit the convenience of rotatory coordinate transformation for eye and head movements, as sketched in Figure 3). In this mapping quandary the brain has an option regarding the design of an analog neural body that is not realizable in a physical body, namely the introduction of a neural fiction – cast in the format of naive realism – of a cyclopean aperture through which the egocenter is interfaced with the visual world from its position inside the head (Merker 2007, p. 73). Such is the normal format of our visual awareness, captured in the drawing by Mach reproduced in Figure 4.

4.3 “Just in time” for the next gaze movement: Brief career of the global best estimate

As we have seen, cortical sensory information is replete with entangled ambiguities in probabilistic form up and down the cortical hierarchies in several modalities. This information must eventually be resolved into a definite conscious percept of the current scene. To do so the information scattered across many cortical areas must undergo a massive operation of multiple constraint satisfaction in which its various contextual and cue dependencies can be brought to bear on one another to settle trade-offs, conflicts and synergies among them, to converge on the brain's best estimate of what “all the sensory facts” in fact amount to. The circumstance that a primary “consumer” of this estimate is the mechanism by which sensory information controls the timing and direction of gaze movements has profound implications for the nature and genesis of the estimate, identified in the foregoing with the contents of our sensory consciousness.

The gaze moves some two to three times a second (the oft cited Figure of 4 to 5 saccades per second is typical of reading, but not of spontaneous viewing

behavior: Rayner 1998). As soon as the gaze moves, the sensory “facts” delivered by sensory arrays change, irrevocably, because the sensory arrays are now differently disposed relative to their sources in the physical world. New items now occupy central vision, delivering new grist for the cortical disambiguation mill. The very next gaze movement is therefore the terminus of the cortical disambiguating operation not only in the sense that it supplies sensory guidance for that gaze movement, but in the sense that *that movement by its very occurrence puts an end to the validity of the premises on which the cortical disambiguating operation that led up to the movement was based*. That leaves only the space between gaze movements – typically a few hundred milliseconds – for an estimate to be extracted from the state of cortical sensory systems, only to become obsolete upon the occurrence of the gaze movement it supports.

The fact that the estimate is rendered obsolete by its own completion suggests that the mechanism producing it be implemented in the form of a temporary storage or buffer. In this representational space cortical probabilities would settle into and register as estimates on a temporary basis only, as sensory stepping stones for the sequential unfolding of orienting behavior. In fact it suggests that buffer contents be erased and replaced with new contents on each gaze movement, a strategy made possible by the fact that the cortical reservoir of probabilistic sensory information is continually available to update buffer contents in a “refresh” type operation. And this exactly because they are *not* – on present assumptions – collapsed into estimates at the cortical level.

It is the sudden change in the premises of cortical disambiguating operations occasioned by a gaze movement that lends utility to such a buffering strategy in the logistics of deriving a sensory best estimate. For the same reason another type of change should likewise trigger a “reset” of the buffer with loss of its previous contents, namely any sudden exogenous change in the sensory situation itself signalled by so called sensory transients (essentially any abrupt and unpredicted change in sensory circumstances). Such transients announce new sensory circumstances, again rendering current premises of cortical disambiguating operations obsolete. Thus they too should occasion the immediate cancellation of the estimate that is in place at the time of the transient and launch a new one incorporating the cortical consequences introduced by the transient.

What is shared by these situations is the utility of abandoning the current contents of the best estimate buffer in favor of an immediate start on a new estimate that captures changed circumstances. This means that even otherwise conspicuous sensory changes introduced during the brief “reset” episodes should typically escape conscious notice. This because the buffer is here assumed to be the locus of conscious content and this content is lost during the reset. It is therefore not

available to the post-reset state of the buffer.² These implications of a hypothetical extracortical global best estimate buffer are supported empirically. Sensory changes introduced at the time of the proposed buffer resets escape conscious detection, as demonstrated in experiments on so called change blindness (Bridgeman & Mayer 1983; Irwin 1991; Rensink et al. 2000; Simons and Ambinder 2005).

The ordinary and routine effect of transients is to attract attention to the specific time and location of a sensory change, and when these specifics are hidden by generalized transients, even when these do not overlap in space with the scene change (“mudsplash” paradigm), the scene change escapes notice, to the consternation of those who are made aware of their failure to notice major changes in a scene. Scene changes that are noticed despite general transients (and, as assumed here, resetting of buffer contents) are those for which focal attention on the relevant location or object comes to bridge the brief reset episode (Rensink et al. 1997).

This has been interpreted as a memory effect supported by focal attention, rescuing the attended content from oblivion (Turatto et al. 2003). The privileged relationship between attention and memory (cf. Merker 2007, p. 77) allows pre- and post-reset *focal* content to be automatically compared, because focal content, unlike buffer contents generally, maintain continuity of memory recording, bringing the change to awareness. Focal attention and its contents thus would be the key factor maintaining continuity of consciousness across frequent resets of buffer content. This contrast between a more capacious and conscious but ephemeral buffer and a likewise conscious attentional focus of limited capacity but tied into memory processes outside the buffer is in good agreement with a range of empirical findings reviewed by Jeremy Wolfe and summarized in his “inattention amnesia” hypothesis (Wolfe 1999; see also Lamme 2004). Moreover, capacity limitations on focal attention would be expected to constrain which aspects of a scene are subsequently available for recall, the degree of constraint being determined by the severity of attentional load. Again, this is born out empirically, in experiments demonstrating so called inattention blindness (Mack, 2003; Most et al. 2005).

2. Nothing prevents pre-reset sensory information from exerting a variety of implicit and priming effects (see, for example, Chun & Nakayama 2000; Silverman & Mack 2006; Yeh & Yang 2009). The arguments in the text pertain exclusively to conscious contents, assumed to be hosted by the conscious awareness system in a temporary, buffered state, leaving open the fate of sensory information feeding those contents. Traces of such activity may persist – without awareness – as priming and implicit effects in other systems, the cortex included. The further question of the extent to which buffer resets are accompanied by phasic concomitants in these other systems (see, for example, Churchland et al. 2010) is an important one, but lies beyond the scope of the present treatment.

The “being in time” of conscious contents, in other words, is rich but ephemeral, barely surviving the present moment, and being consigned to oblivion with each intrusive sensory change or movement of the gaze. The impression of robust continuity of a rich sensory world results from the coherent memory thread trailing focal attention, supplemented and supported by the fact that outside the focus – with its attraction to loci of change – things tend to remain pretty much the same. In the world outside the laboratory sailboats and church steeples do not simply appear or vanish from one moment to the next. Most of the world has not changed in that short spell, and most saccades shift the spotlight of high-resolution perception within the compass of a rather modest change in visual field orientation (this because of the limiting effect of the bony orbit, reinforced by foreshortened peripheral vision under central magnification; see the legend to Figure 3). Each such change nevertheless alters the premises for resolving the entangled ambiguities the cortex evolved to grapple with.

From the cortical “point of view” an extracortical buffer generating estimates in almost punctate serial fashion is a convenience permitting it to keep its probabilistic format intact for continuous and wide-ranging sensory interactions in an unconscious regime of suspended commitments. For the first and foremost consumer of these estimates – the gaze – the buffer is a proximal causal source of its movements. And from the point of view of the creatures endowed with such a buffer it is what makes them conscious, should the perspective sketched here turn out to relate to realities. After these preliminaries regarding design features it is time to consider the manner in which this set of desiderata might be realized in neural terms.

5. Cortex, colliculus and “the other thalamus”

By now a considerable set of demanding design requirements has accumulated that must be accommodated by any realistic neural implementation of a sensory awareness mechanism. It must be capable of a massive multiple constraint satisfaction operation over cortical contents in several modalities.³ It must keep sensory maps in register and the world stable despite relative motion of peripheral sensory

3. As first suggested by Geoffrey Hinton (Hinton 1977), a large class of artificial so called neural networks are in effect performing multiple constraint satisfaction (Rumelhart et al. 1986). The most powerful constraint satisfaction algorithms rely on local exchange of information between variables and constraints (e.g. “survey propagation” – Achlioptas et al. 2005), and are accordingly amenable to parallel implementation (Mézard & Mora 2009). This suggests that the problem we are considering is neurally tractable.

arrays. It must possess decision mechanisms that allow diverse sensory contents to issue in a single spatial estimate pointing the gaze at its very next stepping stone in its serial progress through the world. And it must be tied in to memory mechanisms that allow its focal content to accrete in the longitudinal record of personal history. All this is to be accomplished in the time available between eye movements, i.e. in as little as a couple of hundred milliseconds when demand is high.

Apart from the details discussed under design features, there are two general requirements for implementing this tall order in neural terms. One of these we have already dealt with: the entire operation must issue in a single spatial estimate pointing the gaze at its very next target in the world. To recapitulate: the elaborate sensorimotor machinery of the superior colliculus is the most caudal site exercising sensory control of integrated orienting movements. It is therefore the final site beyond which no further postponement of sensory decision making is possible, and in or before which all sensory ambiguities must be resolved – for the time being. The exquisite layering within the colliculus of sensory maps in multiple modalities within a common premotor framework suggests that at least some of the final decisions are being made within the colliculus itself. This, however, is not to say that all sensory decisions are executed there, because far from all paths between cortex and colliculus are direct ones. Indirect anatomical circuitry connecting the cortex with the premotor neurons of the collicular intermediate lamina ought therefore to be explored for candidate contributions to the brain's conscious awareness buffer.

There are two classes of descending cortical projections to subcortical structures of interest in this regard, namely those of pyramidal cells in cortical layer 6 and cortical layer 5. A second general requirement helps decide between the two. The global best estimate buffer must perform radical operations on cortical *output*, but must studiously avoid affecting cortical *input*. This because primary sensory afference, being the only source of veridical information about states of the world available to the cortex, must not be tampered with on its way up to its cortical destination. Nevertheless there is a need for “tuning” and normalizing operations on this information in its passage up to the cortex. That is, on the one hand the cortex has a need to efferently affect its inputs in informationally “innocuous” ways to make inputs compatible with its own operations and, on the other, it needs to efferently “export” its own activity states for purposes of a best sensory estimate for the control of behavior led by gaze movements.

These two efferent cortical tasks are by no means equivalent nor even necessarily compatible, and I suggest that the well known distinction between layer 6 and layer 5 pyramidal cells is the cortical solution to this dual efferent need. If so, the function of subcortical projections from layer 6 pyramidal cells would be to tune activity “on the way up to” the cortical probability distributions, while layer 5 pyramidal cells serve as “the way out from” those distributions. The synaptology

of the former are appropriately “modulatory” rather than driving (Guillery & Sherman 2002), they are deployed to *all* thalamic nuclei along with other immediately subcortical (“near”) structures such as the claustrum and the basal ganglia, and they collateralize in the thalamic reticular nucleus on their way to their thalamic targets. The reticular nucleus’ inhibitory projection to thalamic nuclei similarly appears to be modulatory by tending to innervate the distal rather than proximal dendrites of thalamic relay neurons (Wang et al. 2001).

Layer 5 descending projections differ from those of layer 6 in all these respects. They target distant brainstem nuclei (Diamond 1979), and typically innervate more proximal subcortical targets by collaterals along the course of these long projections (Guillery 1995). They innervate only some thalamic nuclei (see below), and do not collateralize in the reticular nucleus on passing through it. Their synaptology exhibits high security “driving” type innervation typically concentrated to proximal dendritic locations (Reichova & Sherman 2004). The way “up” to the cortical probability distributions projects sensory afference into a space of far higher dimensionality than that of the projected afferent channel, while the way “down” – ultimately to behavior – involves radical dimensionality reduction from that space onto – ultimately – the numerically modest set of brainstem and spinal motoneurons. From what has been said so far it should be obvious that it is along the downward path that the present search must proceed, that is, along the descending axons of cortical layer 5 pyramidal cells.

A special role in this regard is likely to be played by what we might call “the other thalamus,” schematically demarcated by a heavy black line in Figure 5. It consists of the “higher order” or “association” nuclei of the dorsal thalamus plus a set of adjoining dorsal thalamic nuclei which all project to the basal ganglia, and which collectively we shall call the “extended intralaminar complex.” The latter includes what has been called the “oculomotor thalamus” (for which see Tanaka & Kunitatsu 2011), and consists of the midline, intralaminar, and paralaminar nuclear groups, plus their caudal extension in the limitans, suprageniculate and magnocellular medial geniculate nuclei (see Jones & Powell 1971; Jones 1989). The thalamic higher order or association nuclei themselves make up the bulk of the medial and caudal reaches of the dorsal thalamus, principally as mediodorsal, lateral dorsal, lateral posterior, posterior, and pulvinar nuclei. Taken together the total volume of this “other thalamus” considerably exceeds the aggregate of first order thalamic relay nuclei.

The connective relations of the other thalamus cannot be assimilated to the classical account of the thalamus as a way station relaying sensory and cerebellar information to the cerebral cortex (Sherman 2007). Instead, its two principal parts together provide what amounts to a centrally placed and compact “internal analog map” of the telencephalon as a whole, such that its higher order nuclei

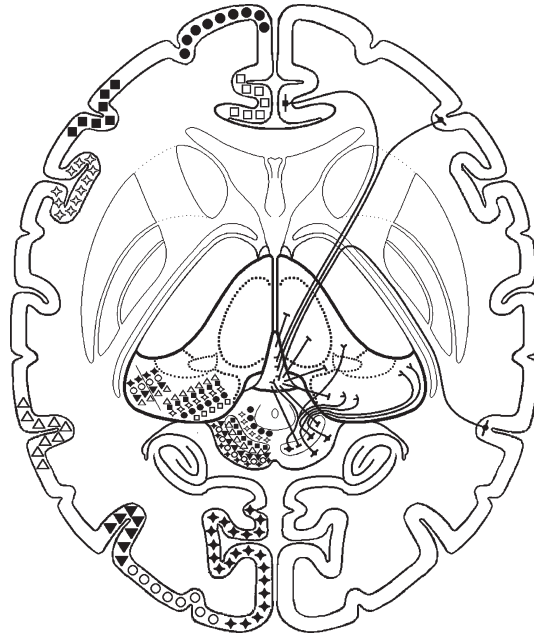


Figure 5. Schematic depiction of connective relations between cortex, thalamus and colliculus discussed in the text. The “other thalamus” is demarcated by a heavy black line, within which dotted outlines mark – in rostrocaudal sequence – the mediodorsal, laterodorsal, lateral posterior and pulvinar nuclei, respectively. The territory surrounding the mediodorsal nucleus is the extended intralaminar complex. A sample of cortical areas are marked with symbols which also mark their highly schematic representation in the pulvinar and superior colliculus. The latter structure has been artificially inserted into the diagram in roughly the position and orientation (“upside down”) in which it would be encountered in horizontal sections of the human brain. The cone-like area at the center of the thalamus is a token representation of the zona incerta, actually located beneath the dorsal thalamus. The thin oblique line through one of the “map stacks” of the pulvinar marks an axis of iso-representation through the aligned maps, each of which is a two-dimensional discoid cut by the plane of the section. Excitatory connections end in a “Y”, inhibitory connections in a “T”. Image by Bjorn Merker licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

are cortically related, while its extended intralaminar complex has as its principal target the other major constituent of the telencephalon, the basal ganglia (Powell & Cowan 1956; Jones 1998), though projecting to the cortex as well (Groenewegen & Berendse 1994). It is this other thalamus – higher order nuclei plus extended intralaminar complex – that is the target of the cortical layer 5 pyramidal cell projection, justifying a collective designation for this large set of thalamic nuclei. The layer 5 projection has “driving” properties, and crucially determines higher order thalamic receptive fields (Bender 1983; Diamond et al. 1992).

There is evidence that the return projection to cortex from higher order thalamic nuclei in a number of cases has driving properties as well (Lee & Sherman 2008). It is as if cortical areas in high hierarchical positions within the cortex, which lie at a greater remove from the primary sensory areas along successive cortico-cortical relay paths, are supplied with a separate source of high-efficacy input through a thalamic nuclear organization entirely separate from the primary thalamic relay nuclei, namely the higher order nuclei. These nuclei themselves receive their input principally from the cortex, and from the full range of its areas (Guillery & Sherman 2002). With such an arrangement, even drastic and irreversible operations on cortical contents can be performed by these nuclei without jeopardizing cortical access to veridical sensory information. Such information is continually being supplied to the primary cortical areas by the sensory relay nuclei of the “first” thalamus, in parallel with the independent operations of the other thalamus. Perhaps, then, those independent operations concern the reduction of cortical probability distributions to definitive but temporary estimates on a running basis.

To begin with, the several nuclei and subdivisions of the other thalamus host massive convergence of afference from the many mapped areas that make up the expanse of the cerebral cortex. This fact by itself would be of little consequence if the convergence simply pooled cortical inputs indiscriminately. Instead it is highly structured, as shown in the best-studied case of the projection of the many visual areas of the macaque cortex onto the visual pulvinar. Not only do these convergent projections maintain retinotopy, but in the pulvinar the resulting visual maps are “stacked” along the third spatial dimension of its solid volume in such a way that cortical maps that present topographic mirror images in the cortical sheet end up “face-to-face” with matched retinotopy in the pulvinar (Shipp 2001, 2003). The result is that any line – or column, if you will – that runs parallel to this third dimension through the visual pulvinar passes through corresponding retinotopic points in all the maps that are successively lined up – sometimes with overlap, and sometimes without – along it.

Thus, neurons occupying a pulvinar iso-representation line will individually reflect the functional specialty of their parent cortical map, but taken together will represent the full cortical range of visual functional specialization. This full range will be represented for each retinal location, in the aggregate covering all of retinal space. The result is that different visual attributes of objects (color, motion etc.) which in the cortical sheet are preferentially represented on separate maps are brought into retinotopic alignment along the iso-representation dimension in the pulvinar. Such an arrangement has the potential of forming a single, unitary visual image space whose moment-to-moment contents incorporate the full scope of multi-areal cortical visual activity that feeds it, and to do so in the form

of a coherent and comprehensive spatial rendition of the current visual scene.⁴ Similar arrangements appear to exist in other parts of the pulvinar: in the dorsal pulvinar (partly coterminous with the medial pulvinar of classical terminology) there is partial overlap between prefrontal and posterior parietal connections in discoid zones interdigitating with zones devoted to connections with other cortical regions (Asanuma et al. 1985).

The principle appears to be that the third dimension available in the pulvinar is used to represent differential functional specialization among cortical maps that share a common format (such as retinotopy), using that format as two of the three thalamic dimensions of representation. The arrangement by and large appears to duplicate an order already established by inter-areal connectivity in the cortex (“replication principle”: Shipp 2003), with the difference that it brings interactions among separate, cortically defined functional specialties within reach of *local* interactions in the compact “map stacks” of the pulvinar. In such a setting local circuitry deployed along and across iso-representation lines within the pulvinar could conceivably provide a mechanism for multiple constraint satisfaction across these cortically driven arrays of aligned specialists by dynamically settling conflicts and synergies among them.

Traditionally the pulvinar – in keeping with thalamic patterns generally – has been regarded as conspicuously lacking in intrinsic connectivity within or between its subdivisions (e.g. Trojanowski & Jacobson 1975; Ogren & Hendrickson 1977). All the more noteworthy, therefore, is the discovery of a novel kind of long-range inhibitory interneuron in the pulvinar (Imura & Rockland 2006). These widely ramifying GABAergic interneurons bear on the present perspective by being concentrated to posterior reaches of the dorsal pulvinar. This is a multimodal region driven by the auditory association cortex, posterior parietal and temporal visual and multimodal areas, as well as with parahippocampal, prefrontal

4. This arrangement may be construed as a solution to the so called perceptual “binding problem,” except that even *cortical* connectivity is such that that problem never arises. Thus, in the visual system, most cortical maps maintain retinotopy (except at the highest end of the hierarchy (Kawato 1997; Lennie 1998)). These maps are connected in topographic and bidirectional fashion in such a way that retinotopic points are matched. Connectively considered the cortex therefore is not a set of maps lying side-by-side but rather a map-stack (neural solid) within which maps lie face-to-face in proper topographic alignment (Merker 2004, p. 564 and Figure 2 therein). This ensures that the contents of maps with different functional specialization are retinotopically aligned and functionally connected along iso-representation lines defined by the trajectory of cortical feedforward and feedback connections (cf. Edelman 2002, p. 130). As explicated by Kawato (1997), such an arrangement obviates the so called “binding problem.”

(including frontal eye fields), orbitofrontal and insular cortices (Gutierrez et al. 2000; Imura & Rockland 2006; Kaas & Lyon 2007; see also Cappe et al. 2009).

It is too early to know whether the unusual scaffolding of long range intrinsic inhibition investing this multimodal domain of the pulvinar enables it to reduce its sophisticated and wideranging cortical afference to a global best estimate of current sensory circumstances. If it does, the preceding arguments predict that its normal operational dynamics must unfold within a full surround (360 degree) egocentrically organized and multimodal representational space, along the lines of the world-stabilizing schematism sketched in Figure 3. The presence of such a space in the caudal pulvinar could be explored by using, say, auditory stimuli for orienting movements to visually absent space,⁵ supplying one possible test for the role conjecturally suggested for it here.

Moreover, the functional precipitate of cortical activity in that neural space should under normal circumstances be structured around an egocenter lodged inside a visual-head-based cyclopean aperture in turn lodged inside and interfaced with a surrounding visual world. This, because for one thing its multiple constraint satisfaction regime presumably must operate by fast dynamics to be neurally tractable (see Footnote 3, and Ma et al. 2006). For another, that regime is fed by cortical, that is, experience- and memory-based, information and thus incorporates the probabilistic priors embedded in cortical operations. Those priors must be shaped by the perpetual framing of visual scenes by the orbital limits of the visual field, yielding cyclopean perspectivalism (viewpoint-based topology) as the most parsimonious global ordering of the perpetual stream of visual stimuli.

By the same token, when primary input or activation patterns of such a dynamic regime are sufficiently altered, by experimental manipulations that include an altered relation between visual body and visual world (Ehrson 2007; Lenggenhager et al. 2007), or by unusual activation states such as might occur during personal emergencies, clinical conditions, or electrical stimulation of the cortex (Blanke et al. 2004) the same constraint satisfaction operation would be expected to arrive at alternate topologies. This occurs in the out-of-body experiences reported in the studies just cited, in which the ego-center no longer lodges inside the visible body. The relative ease with which such conditions are induced

5. The superior colliculus is the only site in the brain so far known to represent visually absent space in a multimodal framework. It is represented in the caudal-most regions of the collicular map, which trigger orienting movements to locations outside the field of view in response to, say, auditory stimuli coming from behind the head. In animals with laterally placed eyes even the visual map extends to virtually 360 degrees. All that is needed to obtain a representation of visually absent space in animals with frontally placed eyes is therefore to retain the more basic collicular sensorimotor map arrangement of animals with lateral eyes.

in the laboratory by appropriate stimulation rules out a structural or innate basis for the habitual topological relationship between ego-center, body and world in experience, and supports a dynamic constraint-settling account of its origin.

Any brain locus hosting the hypothetical global best estimate is of course bound to exhibit these properties, for the reasons just given. So far the pulvinar is no more than a promising connective nexus for such a function. There are, however, empirical indications of a special pulvinar role in sensory awareness. Its longstanding association with attentional dynamics, overt as well as covert (Arend et al. 2008; Bender & Youakim 2001; Hulme et al. 2010; LaBerge & Buchsbaum 1990; Petersen et al. 1985, 1987; Rushmore et al. 2006; Shipp 2004; Snow et al. 2009), fits well with such a role. It is, moreover, the thalamic region most directly implicated in the clinical syndrome of spatial neglect (Rushmore et al. 2006), in humans specifically involving the right pulvinar (Karnath et al. 2002). The neglect syndrome is of special interest from the present point of view, since it comes close to being a pure affliction of sensory awareness. Conjunction errors have also been associated with spatial deficits following pulvinar damage (Ward et al. 2002; Ward & Danziger 2005).

It is also there that – along with a number of cortical areas – hemodynamic responses correlate with perceptual reversals during the viewing of ambiguous figures (Kleinschmidt et al. 1998), a phenomenon which, as noted in an earlier section, occurs in the presence of fixed and unchanging sensory stimulation. And, significantly, a region of the macaque dorsal pulvinar that overlaps with the pulvinar territory populated by long range inhibitory interneurons houses neuronal populations whose responses in a generalized flash suppression paradigm show *higher specificity for stimulus awareness than that exhibited by striate and extrastriate neurons studied with the same method* (Wilke et al. 2009). The same general region of the dorsal pulvinar, on unilateral inactivation, impairs visuomotor operations in contralateral space in ways that resemble disruption of parietal and frontal function in humans (Wilke et al. 2010).⁶ It is also in the general region of the dorsolateral pulvinar of humans that a representation of precise visual localization has been identified by functional imaging methods, remarkably confined to the right pulvinar (Fischer & Whitney 2009). The association of

6. The neuronal populations recorded by Wilke and colleagues (2009) are found at pulvinar levels that overlap with those in which Imura and Rockland identified long range inhibitory interneurons, but also more anteriorly and ventrally (sections b and c in Imura and Rockland Figure 13 come from AP levels corresponding to levels 3 and 4 in Fig. S1 of the Supporting Information of Wilke and colleagues). The same is true of the region of pulvinar inactivation that disrupted visuomotor behavior in monkeys (Wilke et al. 2010): It overlaps with the pulvinar region populated by long range inhibitory interneurons.

right pulvinar damage with unilateral neglect already cited (Karnath et al. 2002) may imply that in humans the right pulvinar contains a complete representation of visual space (both hemifields; see Bisiach & Vallar 2000 for the rationale behind this conjecture). If so, this complete visual field might even function within the 360 degree multimodal spatial surround postulated in the foregoing.

Most cells in the visual pulvinar show fluctuations in responsiveness covarying with the animal's general level of alertness (Bender 1982), and dissociations between pulvinar and cortical electroencephalographic activity in sleep states may indicate that the pulvinar is more closely linked to awareness than is the cortex. It has generally been assumed that the electrical activity of thalamus and cortex evolve in parallel during sleep and waking. By recording through indwelling electrodes in cortex and medial pulvinar Magnin and colleagues have shown that this is not the case during the transition from wakefulness to sleep, in which pulvinar deactivation precedes that of the cortex, nor in rapid eye movement sleep (Magnin et al. 2004; Magnin et al. 2009; Rey et al. 2007). In this latter stage of sleep, in which episodes of dreaming (i.e. awareness) are more frequent than in other sleep stages, the cortex exhibits a continuous activation state resembling waking, while the medial pulvinar goes through stretches of delta-wave deactivation alternating with activation.

Should the pulvinar be part of the brain's global best estimate buffer and hence of sensory awareness, the periods of pulvinar activation during rapid eye movement sleep would presumably mark dreaming episodes, while a continuously activated but unconscious state of the cortex would ensure memory recording of focal conscious content (just as in waking), and hence the possibility of dream recall. A similar conjecture may apply to slow wave sleep as well. Given the multiple similarities between the thalamocortical dynamics of the UP states of slow-wave sleep and the electrophysiological characteristics of continuous wakefulness, why do the former not lead to conscious experience (for details, see Destexhe et al. 2007 who formulated the question)? The slow wave sleep UP states may be a requirement for cortical memory processes with which they are associated during this stage of sleep, while details of state and dynamics specific to the pulvinar may determine awareness. Slow-wave sleep does contain dream content, though not to the same extent as does rapid eye movement sleep, and these episodes might take place during periods of pulvinar activation.

In summary, the other thalamus exhibits a number of structural and functional features recommending it for a role in the reduction of cortical probability distributions to estimates. Whether it does so in the comprehensive fashion required for implementing the brain's global best estimate of sensory circumstances (sensory awareness) postulated in the foregoing remains to be determined, and is by no means a foregone conclusion. It is not clear, for example, whether there exists in the dorsal

pulvinar a map of visual space free of the radical magnification of foveal and central vision that renders V1 an unlikely host of visual awareness, given the absence of such magnification in the metrics of experienced visual space (Wagner 2006; for a telling illustration of V1 magnification, see animation at <http://cns.bu.edu/~eric/>). It is also not clear how the pulvinar territory invested with long-range interneurons might access the premotor circuitry controlling gaze and orienting movements, as it might be expected to do considering the close ties between a global best estimate buffer and gaze and orienting behavior envisaged in the foregoing. The dorsal pulvinar is interconnected with parietal and frontal oculomotor regions, but has no caudally directed premotor output of its own.⁷ It avails little that the extended intralaminar complex with its prominent oculomotor-related output to the basal ganglia (and hence to the superior colliculus: McHaffie et al. 2005) lies “right next door” if the pulvinar does not project to it.

One possibility is that the presumptive pulvinar sensory best estimate serves cortical disambiguation directly, while exerting its effects on the gaze indirectly, via the influence of pulvinar return projections to the cortex, its parietal and frontal gaze-related areas in particular (for which, see Wilke et al. 2010). In such a case, cortical activity related to gaze, orienting and focal attention would, in parallel with the proposed pulvinar constraint satisfaction operation, undergo radical reduction to final estimate form for those purposes in the superior colliculus, as already modelled by Ma and colleagues (Ma et al. 2006). The fact that both pulvinar and colliculus receive their cortical afference from the axons of layer 5 pyramidal cells, and often by collaterals of the very same axons (Guillery 1995), would ensure tight temporal coordination of their activities. The relevant pulvinar territory is, moreover, continually informed of collicular outcomes through the direct projection of the intermediate layers of the colliculus to the dorsal pulvinar (Benevento & Standage 1983; Harting et al. 1980, Kaas & Lyon 2007). This gives the dorsal pulvinar running access to all dimensions of information bearing on a global best estimate of the momentary sensory situation.⁸

7. An exception might be a possible pulvinar projection to the zona incerta, which has been reported (Power et al. 1999) as well as denied (Bartho et al. 2002). For the zona incerta projection to the pulvinar, see Footnote 8.

8. The global sensory best estimate cannot be determined without access to both gaze and vestibular information: buffer contents need to be reset at the time of gaze shifts, and both gaze and vestibular information is required in order to partition sensory motion into the mobile body and stationary visual world domains of their shared representational space. Both gaze and vestibular information is available to the dorsal pulvinar in its cortical and collicular sources of afference, but also through the unique inhibitory innervation of the higher order nuclei by the zona incerta and the anterior pretectal nucleus (Bokor et al. 2005;

The suggestion is, in other words, that the pulvinar does not need a caudally directed output of its own, since the colliculus operates in tandem with it at all times, and on the basis of the very same high-level sources of gaze-relevant cortical information.

Such a two-tiered arrangement, with the superior colliculus at its base, would help explain the role of this midbrain structure in high-level cognitive functions such as *top-down* covert attentional selection for perceptual judgements (Lovejoy & Krauzlis 2010). Contrary to what might be expected, the top-down influence on sensory selection among competing stimuli is not primarily carried by cortico-cortical feedback projections from frontal to posterior cortical areas, but crucially depends on the superior colliculus, as shown by reversible collicular inactivation in the Lovejoy and Krauzlis experiment. This result belongs among a larger body of evidence for collicular involvement in attentional and perceptual selection, overt as well as covert, with spatial neglect and sensory extinction phenomena as symptoms of malfunction (Carello & Krauzlis 2004; Cavanaugh & Wurtz 2004; Fecteau & Munoz 2006; Horowitz & Newsome 1999; Müller et al. 2005; Rushmore et al. 2006; Sprague 1966).

The common denominator of this high-level collicular role is stimulus competition. The just mentioned deficits are most readily produced by disrupting the local collicular balance of excitation and inhibition by interventions anywhere along its afferent pathways (see Rushmore et al. 2006; Shipp 2004), an effect that can be exerted quite locally within the visual field (Pöppel & Richards 1974). Apparently, some final competitive decisions are carried out in the superior colliculus itself, to secondarily be conducted back to the forebrain via the pulvinar and the extended intralaminar complex, making the colliculus an informational bottleneck for some of the brain's highest-level operations. The colliculo-pulvinar

Mitrofanis 2005). These nuclei are not only connected with one another, the superior colliculus, and the intralaminar thalamus (e.g. Benevento & Standage 1983; Berman 1977; Graham & Berman 1981; Terenzi et al. 1995), but to key cortical areas such as the lateral intraparietal (Asanuma et al. 1985) and anterior cingulate cortices (Cadusseau & Roger 1991; Mitrofanis 2005). They receive vestibular projections directly (Horowitz et al. 2005), and in the case of the zona incerta (along with the intralaminar thalamus) also indirectly via the cerebellum (Meng et al. 2007). The zona incerta also features eye movement activity in the form of omnipause responses (Ma et al. 1996). Being inhibitory, a pause in zona incerta input to the higher order nuclei would release these nuclei from its powerful inhibitory influence at the time of eye movements, perhaps as a means to admit the new information that must replace prior buffer content at the time of buffer resets. More generally, the intricate global and midline-straddling intrinsic inhibitory organization and vast external connectivity of the zona incerta is likely to be integral to the functional logic of the higher order nuclei, which it innervates through high-security inhibitory synapses (see Merker 2007, p. 75, for further details).

return route to the cortex presumably plays a role in the memory effects of focal attention discussed in connection with the temporal dynamics of the best estimate buffer, while the colliculo-intralaminar route closes a subcortical basal ganglia loop through the colliculus (McHaffie et al. 2005). The latter, then, might play a role in resetting the putative global best estimate buffer at the time of eye movements. If so, that function easily extends to cover reset activity triggered by sudden sensory or endogenous change as well, given the prominent intralaminar and collicular afference from activation-related neural systems (Dean et al. 1989; Groenewegen & Berendse 1994; Merker 1980; see also Churchland et al. 2010).⁹ The intralaminar projection to cortical layer one is one of several routes by which such reset dynamics could exert their influence at the cortical level as well.

6. Conclusion

The ideas regarding brain mechanisms of consciousness presented in this chapter are heuristic. It proposes that the functional logic of cortical operations imply a need for an extracortical “global best estimate buffer” as a means to complete inherently probabilistic cortical disambiguating operations through a definitive but ephemeral estimate of current sensory circumstances. As explicated more thoroughly in a separate publication (Merker in press), the contents of that buffer are proposed to be conscious, not by virtue of anything being “added” to them in order to “make them conscious”, but by virtue of buffer contents assuming a crucial conformation or topology by dynamically settling into a nested arrangement that places an ego-center in perspectival relation to a neural model of body-world interactions. Finally, the organization of the higher order nuclei of the dorsal thalamus were explored for their potential role in the global best estimate mechanism, with particular attention to the possibility that the caudal reaches of the dorsal pulvinar might host the specifically sensory aspects of the global best estimate buffer, i.e. sensory awareness.

Much remains to be learned about the functional organization and role of the higher order nuclei of the thalamus before we can know whether they can sustain the burden placed upon them by the present conjectural account. I have deliberately skirted the issue of the consequences of destructive lesions of the

9. See Bowker & Morrison (1976) and Steriade et al. (1990) for the bearing of so called ponto-geniculo-occipital waves and eye movement potentials on this issue, and Sherman 2001 and Swadlow & Gusev 2001, for further details.

other thalamus for the conscious state, because the predictions that follow from the present perspective in that regard involve additional issues. They depend, for one thing, on a more detailed account than provided here of the implications for the constitution of the conscious state of the kind of functional division of labor suggested for the tandem relationship between superior colliculus and pulvinar towards the end of the previous section (Cf. Bao & Pöppel 2007; Denny-Brown 1962; Weddell 2004). For another, the role of zona incerta and anterior pretectal nucleus, the two sources of powerful inhibitory innervation of the other thalamus (see Footnote 8), will have to be taken into account in any consideration of the expected consequences of damage to the proposed conscious awareness system. The association between insults to the region of the third ventricle and loss of consciousness is nevertheless worth noting in this regard (see Bogen 1995). This because the higher order thalamic nuclei are even closer to that ventricle, forming its walls as it were, than the intralaminar complex which has been invoked to account for that association (Bogen 1995). Both components, of course, play functional roles in the present exploratory account of the place of the other thalamus in a candidate extracortical conscious awareness system.

While it is too early to know whether the subcortical structures explored in this chapter implement a conscious awareness system, it would seem that determining whether they do or not must involve an assessment of the extent to which their functional organization is capable of accomodating essential characteristics of the organization of consciousness. The attempt to explicate these characteristics, the design features of consciousness if you will, is therefore an essential prerequisite for progress in understanding its neural mechanisms. In sum, the message of this chapter is therefore that sensory consciousness offers rich structural specifics for employment in any attempt to determine whether a given neural system might implement the brain's mechanism of consciousness, thus to host our egocenter, body and world in their phenomenal trajectory through time.

References

- Adelson, E.H. & Bergen, J.R. 1991. The plenoptic function and the elements of early vision. In M.S. Landy & J.A. Movshon (eds), *Computational models of visual processing* (pp. 1–20). Cambridge, MA: MIT Press.
- Alais, D. & Parker, A. 2006. Independent binocular rivalry processes for form and motion. *Neuron*, 52, 911–920.
- Andersen, R.A. & Mountcastle, V.B. 1983. The influence of the angle of gaze upon the excitability of the light-sensitive neurons of the posterior parietal cortex. *Journal of Neuroscience* 3, 532–548.

- Anderson, C.H. & Van Essen, D.C. 1994. Neurobiological computational systems. In J.M. Zurada, R.J. Marks, 2nd & C.J. Robinson (eds), *Computational Intelligence Imitating Life* (pp. 213–222). New York: IEEE Press.
- Andrews, T.J. & Blakemore, C. 1999. Form and motion have independent access to consciousness. *Nature Neuroscience* 2, 405–406.
- Arend, I., Rafal, R. & Ward, R. 2008. Spatial and temporal deficits are regionally dissociable in patients with pulvinar lesions. *Brain* 131, 2140–2152. doi:10.1093/brain/awn135.
- Asanuma, A., Andersen, A. & Cowan, W.M. 1985. The thalamic relations of the caudal inferior parietal lobule and the lateral prefrontal cortex in monkeys: Divergent cortical projections from cell clusters in the medial pulvinar nucleus. *Journal of Comparative Neurology* 241, 357–381.
- Bao, Y. & Pöppel, E. 2007. Two spatially separated attention systems in the visual field: Evidence from inhibition of return. *Cognitive Processing* 8, 37–44. doi:10.1007/s10339-006-0151-x.
- Bar, M. 2004. Visual objects in context. *Nature Reviews Neuroscience* 5, 617–629.
- Barthó, P., Freund, T.F. & Acsády, L. 2002. Selective GABAergic innervation of thalamic nuclei from zona incerta. *European Journal of Neuroscience* 16, 999–1014.
- Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Shadlen, M.N., Latham, P.E. & Pouget, A. 2008. Probabilistic population codes for Bayesian decision making. *Neuron* 60, 1142–1152. doi:10.1016/j.neuron.2008.09.021.
- Bender, D.B. 1982. Receptive-field properties of neurons in the macaque inferior pulvinar. *Journal of Neurophysiology* 48, 1–17.
- Bender, D.B. 1983. Visual activation of neurons in the primate pulvinar depends on cortex but not colliculus. *Brain Research* 279, 258–261.
- Bender, D.B. & Youakim, M. 2001. Effect of attentive fixation in macaque thalamus and cortex. *Journal of Neurophysiology* 85, 219–234.
- Benevento, L.A. & Standage, G.P. 1983. The organization of projections of the retinorecipient and non-retinorecipient nuclei of the pretectal complex and layers of the superior colliculus to the lateral pulvinar and medial pulvinar in the macaque monkey. *Journal Comparative Neurology* 217, 307–336.
- Berman, N. 1977. Connections of the pretectum in the cat. *Journal of Comparative Neurology*, 174, 227–254. doi:10.1002/cne.901740204.
- Bisiach, E. & Vallar, G. 2000. Unilateral neglect in humans. In F. Boller & J. Grafman (eds), *Handbook of Neuropsychology* (pp. 459–502). Amsterdam: Elsevier.
- Blake, R. & Logothetis, N.K. 2002. Visual competition. *Nature Reviews Neuroscience* 3, 13–21.
- Blanke, O., Landis, T., Spinelli, L. & Seeck, M. 2004. Out-of-body experience and autoscopia of neurological origin. *Brain* 127, 243–258. doi:10.1093/brain/awn040.
- Bogen, J.E. 1995. On the neurophysiology of consciousness: 1. An overview. *Consciousness and Cognition* 4, 52–62.
- Bokor, H., Frere, S.G., Eyre, M.D., Slezia, A., Ulbert, I., Luthi, A. & Acsady, L. 2005. Selective GABAergic control of higher-order thalamic relays. *Neuron* 45, 929–940.
- Bowker, R.M. & Morrison, A.R. 1976. The startle reflex and PGO spikes. *Brain Research* 102, 185–190.
- Bridgeman, B. & Mayer, M. 1983. Failure to integrate information from successive fixations. *Bulletin of the Psychonomic Society* 21, 285–286.
- Cadusseau, J. & Roger, M. 1991. Cortical and subcortical connections of the pars compacta of the anterior pretectal nucleus in the rat. *Neuroscience Research* 12, 83–100.

- Cappe, C., Morel, A., Barone, P. & Rouiller, E.M. 2009. The thalamocortical projection systems in primate: An anatomical support for multisensory and sensorimotor interplay. *Cerebral Cortex* 19, 2025–2037. doi:10.1093/cercor/bhn228.
- Carello, C.D. & Krauzlis, R.J. (2004). Manipulating intent: Evidence for a causal role of the superior colliculus in target selection. *Neuron* 43, 575–583.
- Cavanaugh, J. & Wurtz, R.H. 2004. Subcortical modulation of attention counters change blindness. *Journal of Neuroscience* 24, 11236–11243. doi:10.1523/JNEUROSCI.3724-04.2004.
- Chang, S.W.C., Papadimitriou, C. & Snyder, L.H. 2009. Using a compound gain field to compute a reach plan. *Neuron* 64, 744–755.
- Chun, M.M. & Nakayama, K. 2000. On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition* 7, 65–81.
- Churchland, M.M., Yu, B.M., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., Scott, B.B., Bradley, D.C., Smith, M.A., Kohn, A., Movshon, J.A., Armstrong, K.M., Moore, T., Chang, S.W., Snyder, L.H., Lisberger, S.G., Priebe, N.J., Finn, I.M., Ferster, D., Ryu, S.I., Santhanam, G., Sahani, M. & Shenoy, K.V. 2010. Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience* 13, 369–378. doi:10.1038/nn.2501.
- Corbetta, M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Ollinger, J.M., Drury, H.A., Linenweber, M.R., Petersen, S.E., Raichle, M.E., Van Essen, D.C. & Shulman, G.L. 1998. A common network of functional areas for attention and eye movements. *Neuron* 21, 761–773.
- Cox, P.H. 1999. *An initial investigation of the auditory egocenter: Evidence for a “cyclopean ear.”* Doctoral dissertation, North Carolina State University, Raleigh, NC.
- Dean, P., Redgrave, P. & Westby, G.W.M. 1989. Event or emergency? Two response systems in the mammalian superior colliculus. *Trends in Neurosciences*, 12, 137–147. doi:10.1016/0166-2236(89)90052-0.
- Deneve, S. 2008a. Bayesian Spiking Neurons I: Inference. *Neural Computation*, 20, 91–117.
- Deneve, S. 2008b. Bayesian Spiking Neurons II: Learning. *Neural Computation*, 20, 118–145.
- Deneve, S., Latham, P. & Pouget, A. 1999. Reading population codes: A neural implementation of ideal observers. *Nature Neuroscience* 2, 740–745.
- Denny-Brown, D. 1962. The midbrain and motor integration. *Proceedings of the Royal Society of Medicine*, 55, 527–538.
- Diamond, I. 1979. The subdivisions of the neocortex: A proposal to revise the traditional view of sensory, motor, and association areas. *Progress in Psychobiology and Physiological Psychology*, 8, 1–43.
- Diamond, M.E., Armstrong-James, M., Budway, M.J. & Ebner, F.F. 1992. Somatic sensory responses in the rostral sector of the posterior group (POm) and in the ventral posterior medial nucleus (VPM) of the rat thalamus: dependence on the barrel field cortex. *Journal of Comparative Neurology*, 319, 66–84.
- Edelman, S. 2002. Constraining the neural representation of the visual world. *Trends in Cognitive Sciences*, 6, 125–131.
- Edelman, S., Hiles, B.P., Yang, H. & Intrator, N. 2002. Probabilistic principles in unsupervised learning of visual structure: human data and a model. In T.G. Dietterich, S. Becker and Z. Ghahramani (eds), *Advances in Neural Information Processing Systems* (Vol. 14), (pp. 19–26). Cambridge, MA: MIT Press.
- Ehrson, H. 2007. The experimental induction of out-of-body-experiences. *Science*, 317, 1048.
- Fang, F. & He, S. 2005. Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, 8, 1380–1385.

- Fecteau, J.H. & Munoz, D.P. 2006. Saliency, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences*, 10, 382–390.
- Fischer, J. & Whitney, D. 2009. Precise discrimination of object position in the human pulvinar. *Human Brain Mapping*, 30, 101–111. doi:10.1002/hbm.20485.
- Földiák, P. 1993. The ‘ideal homunculus’: Statistical inference from neural population responses. In F.H. Eeckman & J.M. Bower (eds), *Computation and neural systems* (pp. 55–57). Norwell, MA: Kluwer Academic Publishers.
- Freedman, E.G. & Sparks, D.L. 1997. Activity of cells the deeper layers of the superior colliculus of the rhesus monkey: Evidence for a gaze displacement command. *Journal of Neurophysiol*, 78, 1669–1690.
- Freeman, W.T. 1994. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368, 542–545.
- Gamper, E. 1926. Bau und Leistungen eines menschlichen Mittelhirnwesen (Arhinencephalie mit Encephalocle). *Zeitschrift für die gesamte Neurologie und Psychiatrie*, 102, 154–235, and 104, 49–120.
- Girosi, F., Jones, M. & Poggio, T. 1995. Regularization theory and neural networks architectures. *Neural Computation*, 7, 219–269. doi:10.1162/neco.1995.7.2.219.
- Gorbet, D.J. & Sergio, L.E. 2009. The behavioural consequences of dissociating the spatial directions of eye and arm movements. *Brain Research*, 1284, 77–88.
- Graham, J. & Berman, N. (1981). Origins of the pretectal and tectal projections to the central lateral nucleus in the cat. *Neuroscience Letters*, 26, 209–214.
- Green, A. 2003. The science and philosophy of consciousness. Webpage <http://www.users.globalnet.co.uk/~lka/conz.htm>
- Groenewegen, H.J. & Berendse, H.W. 1994 The specificity of the nonspecific midline and intralaminar thalamic nuclei. *Trends in Neurosciences*, 17, 52–57. doi:10.1016/0166-2236(94)90074-4.
- Guillery, R.W. 1995. Anatomical evidence concerning the role of the thalamus in corticocortical communication: A brief review. *Journal of Anatomy*, 187, 583–592.
- Guillery, R.W. & Sherman, S.M. 2002. Thalamic relay functions and their role in corticocortical communication: Generalizations from the visual system. *Neuron*, 33, 163–175.
- Guitton, D., Munoz, D.P. & Galiana, H.L. 1990. Gaze control in the cat: Studies and modeling of the coupling between orienting eye and head movements in different behavioral tasks. *Journal of Neurophysiology*, 64, 509–531.
- Gu, Y., Angelaki, D.E. & DeAngelis, G.C. 2008. Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience*, 11, 1201–1210.
- Gutierrez, C., Cola, M.G., Seltzer, B. and Cusick, C. 2000. Neurochemical and connectional organization of the dorsal pulvinar complex in monkeys. *Journal of Comparative Neurology*, 419, 61–86.
- Harding, D.E. 1961. *On having no head: Zen and the re-discovery of the obvious*. London: Arkana (Penguin).
- Harting, J.K., Huerta, M.F., Frankfurter, A.J., Strominger, N.L. & Royce, G.J. 1980. Ascending pathways from the monkey superior colliculus: An autoradiographic study. *Journal of Comparative Neurology*, 192, 853–882.
- Hatfield, G. 2003. Representation and constraints: The inverse problem and the structure of visual space. *Acta Psychologica*, 114, 355–378.
- Hayhoe, M. & Ballard, D. 2005. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9, 188–194.

- Helmholtz, H. Von. 1867. *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.
- Hering, E. 1879/1942. *Spatial sense and movements of the eye*, C.A. Radde (transl). American Academy of Optometry. (Original published in 1879).
- Hinton, G.E. (1977). Relaxation and its role in vision. Ph.D. thesis, University of Edinburgh.
- Hinton, G.E. & Brown, D.A. 2000. Spiking Boltzmann machines. *Advances in Neural Information Processing*, 12, 122–128.
- Hinton, G.E. & Sejnowski, T. 1983. Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Washington, D. C., June, 1983*, 448–453.
- Horowitz, S.S., Blanchard, J. & Morin, L.P. 2005. Medial vestibular connections with the hypocretin (orexin) system. *Journal of Comparative Neurology*, 487, 127–146.
- Horowitz, G.D. & Newsome, W.T. 1999. Separate signals for target selection and movement specification in the superior colliculus. *Science*, 284, 1158–1161.
- Howard, I.P. & Templeton, W.B. 1966. *Human spatial orientation*. New York: Wiley.
- Howe, C.Q. & Purves, D. 2005. *Perceiving geometry: Geometrical illusions explained by natural scene statistics*. New York: Springer.
- Howe, C.Q., Lotto, R.B. & Purves, D. 2006. Empirical approaches to understanding visual perception. *Journal of Theoretical Biology*, 241, 866–875.
- Imura, K. & Rockland, K.S. 2006. Long-range interneurons within the medial pulvinar nucleus of macaque monkeys. *Journal of Comparative Neurology*, 498, 649–666.
- Irwin, D.E. 1991. Information integration across saccadic eye movements. *Cognitive Psychology*, 23, 420–456.
- James, W. 1890. *Principles of psychology*, London: Macmillan.
- Jiang, Y. & He, S. 2006. Cortical responses to invisible faces: Dissociating subsystems for facial-information processing. *Current Biology*, 16, 2023–2029.
- Johansson, R.S., Westling, G., Bäckström, A. & Flanagan, J.R. 2001. Eye-hand coordination in object manipulation. *Journal of Neuroscience*, 21, 6917–6932.
- Jones, E.G. & Powell, T.P.S. 1971. An analysis of the posterior group of thalamic nuclei on the basis of its afferent connections. *Journal of Comparative Neurology*, 143, 185–215. doi:10.1002/cne.901430204.
- Jones, E.G. 1989. Defining the thalamic intralaminar nuclei in the monkey. In G. Gainotti, M. Bentevoglio, P. Bergonzi & F.M. Ferro (eds), *Neurologia e Scienze di Base. Scritti in Onore di Giorgio Macchi*, 161–193. Milan: Università Cattolica del Sacro Cuore.
- Jones, E.G. 1998. A new view of specific and nonspecific thalamocortical connections. *Advances in Neurology*, 77, 49–73.
- Julesz, B. 1971. *Foundations of Cyclopean Perception*. Chicago: The University of Chicago Press.
- Kaas, J.H. & Lyon, D.C. 2007. Pulvinar contributions to the dorsal and ventral streams of visual processing in primates. *Brain Research Reviews*, 55, 285–296.
- Karnath, H.-O., Himmelbach, M. & Rorden, C. 2002. The subcortical anatomy of human spatial neglect: Putamen, caudate nucleus and pulvinar. *Brain*, 125, 350–360.
- Kawato, M. 1997. Bi-directional theory approach to consciousness. In M. Ito, Y. Miyashita & E.T. Rolls (eds), *Cognition, computation and consciousness*, 233–248. Oxford: Clarendon Press.
- Kleinschmidt, A., Büchel, C., Zeki, S. & Frackowiak, R.S.J. 1998. Human brain activity during spontaneously reversing perception of ambiguous figures. *Proceedings of the Royal Society (London) B* 265, 2427–2433. doi:10.1098/rspb.1998.0594.
- Knill, D. & Richards, W. (eds). 1996. *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

- Knill, D.C. & Saunders, J.A. 2003. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, 43, 2539–2558.
- Körding, K. & Wolpert, D. 2004. Bayesian integration in sensorimotor learning. *Nature*, 427, 244–247.
- Kornmeier, J. & Bach, M. 2009. Object perception: When our brain is impressed but we do not notice it. *Journal of Vision*, 9, 1–10. doi:10.1167/9.1.7.
- LaBerge, D. & Buchsbaum, M.S. (1990). Positron emission tomographic measurements of pulvinar activity during an attention task. *Journal of Neuroscience*, 10, 613–619.
- Lamme, V.A.F. 2004. Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks*, 17, 861–872.
- Lee, D.C., Hebert, M. & Kanade, T. 2009. Geometric reasoning for single image structure recovery. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, Florida, June, 2009, 2136–2143.
- Lee, C.C. & Sherman, S.M. 2008. Synaptic properties of thalamic and intracortical inputs to layer 4 of the first- and higher-order cortical areas in the auditory and somatosensory systems. *Journal of Neurophysiology*, 100, 317–326. doi:10.1152/jn.90391.2008.
- Lehar, S. 2002. *The world in your head: A gestalt view of the mechanism of conscious experience*. Mahwah, NJ: Erlbaum.
- Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. 2007. Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317, 1096–1099.
- Leopold, D.A. & Logothetis, N.K. 1999. Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, 3, 254–264.
- Leopold, D.A., Maier, A., Wilke, M. Logothetis, N.K. 2005. Binocular rivalry and the illusion of monocular vision. In D. Alais & R. Blake (eds), *Binocular Rivalry*, 231–258. Cambridge, MA: MIT Press.
- Lennie, P. 1998. Single units and visual cortical organization. *Perception*, 27, 889–935.
- Li, Y., Pizlo, Z. & Steinman, R.M. 2009. A computational model that recovers the 3D shape of an object from a single 2D retinal representation. *Vision Research*, 49, 979–991.
- Lovejoy, L.P. & Krauzlis, R.J. 2010. Inactivation of primate superior colliculus impairs covert selection of signals for perceptual judgments. *Nature Neuroscience*, 13, 261–267.
- Ma, T.P. 1996. Saccade-related omnivectoral pause neurons in the primate zona incerta. *Neuroreport*, 7, 2713–2716.
- Ma, W.J. 2010. Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, 50, 2308–2319. doi:10.1016/j.visres.2010.08.035.
- Ma, W.J., Beck, J.M., Latham, P.E. & Pouget, A. 2006. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9, 1432–1438. doi:10.1038/nn1790.
- Mach, E. 1897. *Contributions to the analysis of the sensations*. La Salle, Illinois: Open Court.
- Mack, A. 2003. Inattention blindness: Looking without seeing. *Current Directions in Psychological Science*, 12, 180–184.
- Magnin, M., Bastuji, H., Garcia-Larrea, L. & Mauguière, F. 2004. Human thalamic medial pulvinar nucleus is not activated during paradoxical sleep. *Cerebral Cortex*, 14, 858–862. doi:10.1093/cercor/bhh044.
- Magnin, M., Rey, M., Bastuji, H., Guillemant, P., Mauguière, F. & Garcia-Larrea, L. 2009. Thalamic deactivation at sleep onset precedes that of the cerebral cortex in humans. *Proceedings of the National Academy of Sciences (USA)*, 107, 3829–3833.
- Marigold, D.S. 2008. Role of peripheral visual cues in online visual guidance of locomotion. *Exercise and Sport Sciences Reviews*, 36, 145–151.

- Marigold, D.S. & Patla, A.E. 2007. Gaze fixation patterns for negotiating complex ground terrain. *Neuroscience*, 144, 302–313.
- Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society London B*, 275, 483–519. doi:10.1098/rstb.1976.0090.
- Marroquin, J., Mitter, S. & Poggio, T. 1987. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82, 76–89.
- McHaffie, J.G., Stanford, T.R., Stein, B.E., Coizet, V. & Redgrave, P. 2005. Subcortical loops through the basal ganglia. *Trends in Neurosciences*, 28, 401–407.
- Meng, H., May, P.J., Dickman, J.D. & Angelaki, D.E. 2007. Vestibular signals in primate thalamus: Properties and origins. *Journal of Neuroscience*, 27, 13590–13602.
- Merker, B. (1980). The sentinel hypothesis: A role for the mammalian superior colliculus. Doctoral dissertation, M.I.T. Department of Psychology and Brain Science, Cambridge, MA.
- Merker, B. 2004. Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex*, 40, 559–576. Preprint at: <http://cogprints.org/6456/>
- Merker, B. 2007. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. Target article, peer commentary and author's response. *Behavioral and Brain Sciences*, 30, 63–134.
- Merker, B. In press. Nested ontology and causal options: A paradigm for consciousness. In A. Pereira Jr. & D. Lehmann (eds), *The unity of mind, brain and world: Current perspectives on a science of consciousness*. Cambridge: Cambridge University Press.
- Mezard, M. & Mora, T. 2009. Constraint satisfaction and neural networks: A statistical physics perspective. *Journal of Physiology Paris*, 103, 107–113.
- Mitrofanis, J. 2005. Some certainty for the “zone of uncertainty”? Exploring the function of the zona incerta. *Neuroscience*, 130, 1–15.
- Most, S., Scholl, B., Clifford, E. & Simons, D. 2005. What you see is what you set: Sustained inattentive blindness and the capture of awareness. *Psychological Review*, 112, 217–242.
- Müller, J.R., Philastides, M.G. & Newsome, W.T. 2005. Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences USA*, 102, 524–529. doi:10.1073/pnas.0408311101.
- Munoz, D.P., Guitton, D. & Pélisson, D. 1991. Control of orienting gaze shifts by the tectoreticulospinal system in the the head-free cat. III. Spatiotemporal characteristics of phasic motor discharges. *Journal of Neurophysiology*, 66, 1642–1666.
- Neelon, M.F., Brungart, D.S. & Simpson, B.D. 2004. The isoazimuthal perception of sounds across distance: A preliminary investigation into the location of the audio egocenter. *Journal of Neuroscience*, 24, 7640–7647.
- Ogren, M.P. & Hendrickson, A.E. 1977. The distribution of pulvinar terminals in visual areas 17 and 18 of the monkey. *Brain Research*, 137, 343–350.
- Petersen, S.E., Robinson, D.L. & Keys, W. 1985. Pulvinar nuclei of the behaving monkey: Visual responses and their modulation. *Journal of Neurophysiology*, 54, 867–886.
- Petersen, S.E., Robinson, D.L. & Morris, J.D. 1987. Contributions of the pulvinar to visual spatial attention. *Neuropsychologia*, 25, 97–105.
- Philipona, D., O'Regan, J.K. & Nadal, J.P. 2003. Is there something out there? Inferring space from sensorimotor dependencies. *Neural Computation*, 15, 2029–2049.
- Philipona, D., O'Regan, J.K., Nadal, J.P. & Coenen, O.J.-M.D. 2004. Perception of the structure of the physical world using multimodal unknown sensors and effectors. *Advances in Neural Information Processing Systems*, 16, 945–952.
- Pizlo, Z. 2001. Perception viewed as an inverse problem. *Vision Research*, 41, 3145–3161 doi:10.1016/S0042-6989(01)00173-0.

- Pizlo, Z., Li, Y. & Steinman, R.M. 2008. Binocular disparity only comes into play when everything else fails: A finding with broader implications than one might suppose. *Spatial Vision*, 21, 495–508.
- Poggio, T. & Koch, C. 1985. Ill-posed problems in early vision: From computational theory to analogue networks. *Proceedings of the Royal Society of London. Series B*, 226, 303–323.
- Pollen, D.A. 2011. On the emergence of primary visual perception. *Cerebral Cortex*, 21, 1941–1953. doi:10.1093/cercor/bhq285.
- Powell, T.P.S. & Cowan, W.M. 1956. A study of thalamo-striate relations in the monkey. *Brain*, 79, 364–389.
- Power, B.D., Kolmac, C.I. & Mitrofanis, J. 1999. Evidence for a large projection from the zona incerta to the dorsal thalamus. *The Journal of Comparative Neurology*, 404, 554–565.
- Pouget, A., Dayan, P. & Zemel, R.S. 2003. Inference and computation with population codes. *Annual Review of Neuroscience*, 26, 381–410. doi:10.1146/annurev.neuro.26.041002.131112.
- Pöppel, E. & Richards, W. 1974. Light sensitivity in cortical scotomata contralateral to small islands of blindness. *Experimental Brain Research*, 21, 125–130.
- Purves, D., Lotto, R.B., Williams, S.M., Nundy, S. & Yang, Z. 2001. Why we see things the way we do: Evidence for a wholly empirical strategy of vision. *Philosophical Transactions of the Royal Society London (B)*, 365, 285–297. doi:10.1098/rstb.2000.0772.
- Rao, R.P.N. & Ballard, D.M. 1997. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9, 805–847.
- Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 372–422. doi:10.1037/0033-2909.124.3.372.
- Reichova, I. & Sherman, S.M. 2004. Somatosensory corticothalamic projections: Distinguishing drivers from modulators. *Journal of Neurophysiology*, 92, 2185–2197. doi:10.1152/jn.00322.2004.
- Rensink, R.A., O'Regan, J.C. & Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8, 368–373.
- Rensink, R.A., O'Regan, J.C. & Clark, J.J. 2000. On the failure to detect changes in scenes across brief interruptions. *Visual Cognition*, 7, 127–145.
- Rey, M., Bastuji, H., Garcia-Larrea, L., Guillemant, P., Mauguière, F. and Magnin, M. 2007. Human thalamic and cortical activities assessed by dimension of activation and spectral edge frequency during sleep wake cycles. *Sleep*, 30, 907–912.
- Rizzolatti, G., Riggio, L., Dascola, I. & Umiltà, C. 1987. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A), 31–40.
- Roelofs, C.O. 1959. Considerations on the visual egocenter. *Acta Psychologica*, 16, 226–34.
- Roger, A.S. & Schwartz, E.L. 1999. Design characteristics for a space-variant visual sensor with complex-logarithmic geometry. *Proceedings 10th International Conference on Pattern Recognition* (1990), vol. 2, 278–285. IEEE Computer Society Press. doi:10.1109/ICPR.1990.119370.
- Rumelhart, D.E., Smolensky, P., McClelland, J.L. & Hinton, G.E. (1986). Schemata and sequential thought processes in PDP models. In D.E. Rumelhart & J.L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructure*, vol. 2: psychological and biological models (pp. 7–57). Computational models of cognition and perception series. Cambridge MA: The MIT Press.
- Rushmore, R.J., Valero-Cabre, A., Lomber, S.G., Hilgetag, C.C. & Payne, B.R. 2006. Functional circuitry underlying visual neglect. *Brain*, 129, 1803–1821.

- Schacter, D.L. 1989. On the relations between memory and consciousness: Dissociable interactions and conscious experience. In H.L. Roediger, III & F.I.M. Craik (eds), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving* (pp. 356–390). Mahwah, NJ: Lawrence Erlbaum Associates.
- Schacter, D.L. 1990. Toward a cognitive neuropsychology of awareness: Implicit knowledge and anosognosia. *Journal of Clinical and Experimental Psychology*, 12, 155–178. doi:10.1080/01688639008400962.
- Schwartz, O., Hsu, A. & Dayan, P. 2007. Space and time in visual context. *Nature Reviews Neuroscience*, 8, 522–535.
- Sherman, S.M. 2001. A wake-up call from the thalamus. *Nature Neuroscience*, 4, 344–346.
- Sherman, S.M. 2007. The thalamus is more than just a relay. *Current Opinion in Neurobiology*, 17, 417–422. doi:10.1016/j.conb.2007.07.003.
- Shipp, S. 2001. Corticopulvinar connections of areas V5, V4, and V3 in the Macaque monkey: A dual model of retinal and cortical topographies. *Journal of Comparative Neurology*, 439, 469–490.
- Shipp, S. 2003. The functional logic of cortico-pulvinar connections. *Philosophical Transactions of the Royal Society London B*, 358, 1605–1624.
- Shipp, S. 2004. The brain circuitry of attention. *Trends in Cognitive Sciences*, 8, 223–230.
- Silverman, M.E. & Mack, A. 2006. Change blindness and priming: When it does and does not occur. *Consciousness and Cognition*, 15, 409–422.
- Simons, D.J. & Ambinder, M.S. 2005. Change blindness. Theory and consequences. *Current Directions in Psychological Science*, 14, 44–48.
- Snow, J.C., Allen, H.A., Rafal, R.D. & Humphreys, G.W. 2009. Impaired attentional selection following lesions to human pulvinar: Evidence for homology between human and monkey. *Proceedings of the National Academy of Sciences USA*, 106, 4054–4059.
- Sokolov, E.N. 1963. Higher Nervous Functions: The Orienting Reflex. *Annual Review of Physiology*, 25, 545–580. doi:10.1146/annurev.ph.25.030163.002553.
- Sprague, J.M. 1966. Interaction of cortex and superior colliculus in mediation of visually guided behavior in the cat. *Science*, 153, 1544–1547.
- Stephen, D.G. & Mirman, D. 2010. Interactions dominate the dynamics of visual cognition. *Cognition*, 115, 154–165.
- Steriade, M., Paré, D., Datta, S., Oakson, G. & Curró Dossi, R. 1990. Different cellular types in mesopontine cholinergic nuclei related to Ponto-Geniculo-Occipital waves. *Journal of Neuroscience*, 10, 2560–2579.
- Sterzer, P., Haynes, J.-D. & Rees, G. 2008. Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. *Journal of Vision*, 8(15)10. doi:10.1167/8.15.10.
- Swadlow, H.A. & Gusev, A.G. 2001. The impact of “bursting” thalamic impulses at a neocortical synapse. *Nature Neuroscience*, 4, 402–408.
- Tanaka, M. & Kunimatsu, J. 2011. Contribution of the central thalamus to the generation of volitional saccades. *European Journal of Neuroscience*, 33, 2046–2057.
- Terenzi, M.G., Zagon, A. & Roberts, M.H.T. 1995. Efferent connections from the anterior pretectal nucleus to the diencephalon and mesencephalon in the rat. *Brain Research*, 701, 183–191. doi:10.1016/0006-8993(95)01001-8.
- Tononi, G. & Edelman, G.M. 1998. Consciousness and complexity. *Science*, 282, 1846–1851.
- Treuhub, A. 1991. *The cognitive brain*. Cambridge, MA: MIT Press.
- Trevarthen, C. 1968. Two mechanisms of vision in primates. *Psychologische Forschung*, 31, 229–337.

- Trojanowski, J.Q. & Jacobson, S. (1975). Peroxidase labeled subcortical afferents to pulvinar in rhesus monkey. *Brain Research*, 97, 144–150.
- Turatto, M., Bettella, S., Umiltà, C. & Bridgeman, B. 2003. Perceptual conditions necessary to induce change blindness. *Visual Cognition*, 10, 233–255.
- van der Kooij, K. & te Pas, S.F. 2010. Shape contrast: A global mechanism? *Vision Research*, 50, 2455–2459.
- van Rossum, M.C.W., Turrigiano, G.G. & Nelson, S.B. 2002. Fast propagation of firing rates through layered networks of noisy neurons. *Journal of Neuroscience*, 22, 1956–1966.
- Velmans, M. 1991. Is human information processing conscious? *Behavioral and Brain Sciences*, 14, 651–726.
- Wagner, M. 2006. *The geometries of visual space*. Mahwah, NJ: Erlbaum.
- Wang, S., Bickford, M.E., Van Horn, S.C., Erisir, A., Godwin, D.W. & Sherman, S.M. 2001. Synaptic targets of thalamic reticular nucleus terminals in the visual thalamus of the cat. *Journal of Comparative Neurology*, 440, 321–341.
- Ward, R. & Danziger, S. 2005. Selective attention and response control following damage to the human pulvinar. In G.W. Humphreys & M.J. Riddoch (eds), *Attention in action: Advances from cognitive neuroscience*, 325–350. Hove, East Sussex: Psychology Press.
- Ward, R., Danziger, S., Owen, V. & Rafal, R. 2002. Deficits in spatial coding and feature binding following damage to spatiotopic maps in the human pulvinar. *Nature Neuroscience*, 5, 99–100.
- Watson, A.B. 1987. Efficiency of a model human image code. *Journal of the Optical Society of America A*, 4, 2401–2417.
- Weddell, R.A. 2004. Subcortical modulation of spatial attention including evidence that the Sprague effect extends to man. *Brain and Cognition*, 55, 497–506.
- Wilke, M., Mueller, K.M. & Leopold, D.A. 2009. Neural activity in the visual thalamus reflects perceptual suppression. *Proceedings of the National Academy of Sciences USA*, 106, 9465–9470.
- Wilke, M., Turchi, J., Smith, K., Mishkin, M. & Leopold, D.A. 2010. Pulvinar inactivation disrupts selection of movement plans. *Journal of Neuroscience*, 30, 8650–8659.
- Witkin, A.P. 1981. Recovering surface shape and orientation from texture. *Artificial Intelligence*, 17, 17–45.
- Wolfe, J.M. 1999. Inattentional amnesia. In V. Coltheart (ed), *Fleeting memories* (pp. 71–94). Cambridge, MA: MIT Press.
- Yeh, Y. & Yang, C. 2009. Is a pre-change object representation weakened under correct detection of a change? *Consciousness and Cognition*, 18, 91–102.
- Zemel, R.S., Dayan, P. & Pouget, A. 1998. Probabilistic interpretation of population codes. *Neural Computation*, 10, 403–430.
- Zettl, J.L., Holbeche, A., McIlroy, W.E. & Maki, B.E. 2005. Redirection of gaze and switching of attention during rapid stepping reactions evoked by unpredictable postural perturbation. *Experimental Brain Research*, 165, 392–401.

Being in time

Shimon Edelman & Tomer Fekete*

Cornell University, Ithaca, NY, USA / Simtat hanarkis 1, Zichron Yaakov, Israel

1. A few more constraints on theories of phenomenal experience

Computational theories of phenomenal experience are steadily emerging as the only serious game in town in consciousness research (e.g. Baars 1988; Dennett 1991; O'Brien & Opie 1999; McDermott 2001; Sloman & Chrisley 2003; Metzinger 2003; Merker 2007; Tononi 2008; Shanahan 2010). In a recent paper, we (Fekete & Edelman 2011) have posited a set of constraints, dictated by the choice of computational framework and by implementational considerations, which any such theory must arguably fulfill. In the present short chapter, we amplify and extend the discussion of those theoretical considerations that have to do with *time* – a critically important concern, given the basic and inherent temporality of phenomenal experience (James 1890, p. 608; see Lloyd this volume, for a review). Specifically, we focus on the interaction between the time-related computational requirements of tractability and timeliness on the one hand, and, on the other hand, another basic characteristic of phenomenality: its *autonomy*.

1.1 Autonomy

Any theory that purports to explain how phenomenal experience arises in the system in question, be it the brain or some other kind of computing environment, must do so in terms that are meaningful within the system itself. Simply put, because my experience of the world cannot be up to an outside observer, it may not be explained by appealing to theoretical constructs that can only be fleshed out from outside my brain. In particular, a theory of experience may not

* Shimon Edelman is with the Department of Psychology, Cornell University, Ithaca, NY 14853-7601, USA. Tomer Fekete is with the Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY 11794-5281, USA.

invoke any function of brain activity that is not computed by the brain itself. This is the sense in which theories of phenomenal experience must be *intrinsic* (Fekete 2010; Fekete & Edelman 2011). We shall have more to say about this notion in Section 2.

1.2 Timeliness

Insofar as phenomenal experience is not a frozen state but a process that unfolds in real time,¹ a theory that aims to explain it may not settle for too loose a connection between the timing of experience and the timing of processes that purportedly constitute it. Furthermore, because the real-time experience is distinct from whatever higher-order scrutiny it may be subjected to *post hoc*, a theory that aims to explain it may not rely on counterfactuals. That is, such a theory must deal in what the brain is doing, or what it has been doing over an appropriately short time window leading up to the present moment – not in what the brain could have been doing, were some of its parameters set to different values.

1.3 Computational tractability

Theories of phenomenal experience that are computational are subject to another, rather obvious, constraint: the computations to which experience is reduced must be *tractable*, given the physical limitations of the implementing system such as the brain. The notion of tractability that applies here is the standard one from computer science (Garey & Johnson 1979), according to which a computational “explanation” is ruled out not only if it reduces to a problem that cannot be solved in principle (e.g. the Halting Problem), but also if it requires time that grows exponentially with a relevant measure of the problem size (say, the number of neurons in an assembly that needs to reach a consensus). Moreover, as per the timeliness constraint stated above, even if the combinatorial complexity of the approach in question is benign, it may not involve waiting times that can grow too long. Thus, if a transition between two experiential states is to be explained by a computation performed on a certain data set, the brain must be capable of completing that computation within an appropriate time frame.

1. In systems engineering, a process is said to respond to events in real time if any delay between the event and the response to it is bounded from above by a fixed bound.

2. Implications of the autonomy of experience

The notion that a theory of experience should be intrinsic because experience is autonomous turns out to be surprisingly tricky to translate into practice. If left unqualified, it is liable to become a wrecking ball that will leave not a single stone standing in any reasonably detailed theoretical landscape. Consider this: all computational theories of experience – indeed, theories of any brain function – are formulated in terms of brain states. The notion of a brain state (e.g. Churchland 1987; Churchland & Sejnowski 1992; Brown 2006; Spivey 2006) is underwritten by the assumption, which typically goes without saying, that the instantaneous state of some or all of the brain neurons is a viable theoretical primitive. As a methodological position, it is shared, for instance, by all theories that employ the tools of linear algebra, including vector spaces and matrix and tensor calculus. As a conceptual assumption, it underlies the idea that stimulus representation, or even phenomenal experience (Smart 2004), may be identified with a point in a multidimensional space, each dimension being the instantaneous state of one of the brain's neurons (or something to that effect). And yet, a multidimensional brain state is an abstraction that, on a strictly intrinsic stance, may be branded a fiction.

For a theorist in search of a certifiably intrinsic account, the difficulty with the concept of a multidimensional brain state stems from such a state being necessarily distributed over space. Insofar as the speed of information propagation in the brain, as elsewhere in the universe, cannot exceed the speed of light in vacuum,² the simultaneous collective state of a spatially distributed system is ill-defined. Worse, it is ill-defined both intrinsically and extrinsically: neither the system's own elements nor any external observer³ can have guaranteed access to the collective state. How then can the concept of brain state have any explanatory value?

The true extent of the conceptual mess into which we just maneuvered ourselves becomes apparent once we realize that the above line of argumentation lends itself to being extended *ad absurdum*: it can be made to undermine not just brain science but all of physics. With this realization comes a glimpse of a way out of the conundrum. Should we be concerned that a physical system cannot know its own instantaneous description in terms of its phase or state space? Not really: the laws of physics couched in terms of such abstractions can be both predictively effective and reductively explanatory (as when a hitherto mysterious concept,

2. This, of course, is one of the postulates of the Special Relativity Theory (Einstein 1905).

3. Unless it is both punctate (i.e. of zero spatial extent) and equidistant from all the system's components.

such as temperature, is reduced in thermodynamics to a collective property of the system); no less and no more should be expected from a good theory. And what of the concern voiced in the previous paragraph? It is rooted in a misconception – namely, in the expectation that for the system to function as a whole, its components must somehow “know themselves” in a global sense, over and above being individually subject to the laws of physics.

3. Time is of the essence

Having briefly skirted a meta-theoretical disaster, let us now return to the conceptual home territory of the cognitive and neural sciences: state spaces. What’s good for the goose, should be good for the gander: given that the concept of an instantaneous system state is, after all, theoretically respectable, and given its use in physics, perhaps it will prove equally useful in neuroscience, and in particular in the neuroscience of phenomenal experience. Alas, this is not to be. The instantaneous state approach is missing a critical conceptual component, without which it remains unfit for use even in theories of goal-driven cognitive representation, let alone of phenomenal experience. The critical component in question is *dynamics* – a formalization of the dependence of the system’s future on its past.

In physical theories, this dependence is expressed through the use of differential equations in which time derivatives of the system’s state variables appear alongside the variables themselves. Without resorting to time-dependent quantities, physics and other basic sciences would be unable to explain the behavior of the world as we know it. In cognitive science too, casting representational theories in terms of instantaneous brain states is problematic from the explanatory standpoint because any such state, taken in isolation from its past, says nearly nothing about the system’s dynamics (Fekete & Edelman 2011). To appreciate this fact, consider the following observation: if the only thing that matters about the present state of a representational system is the point it occupies at this instant of time in the space of its possible states, then any other system that happens to be isomorphic to it must be representationally identical to it. That, in turn, implies that a sack with as many black pebbles as you have active neurons at this instant and as many white pebbles as you have inactive neurons is in same representational (or even experiential) state as you are.

Unless you embrace panpsychism (a strictly last-resort option, we believe), you will find this proposition absurd. Some consider this to be a proof that computation, *qua* manipulation of representations, is always in the eye of the beholder (that is, not intrinsic to the system, in our terminology), and therefore that

computational theories of cognition are out (e.g. Searle 1980). The computational approach to cognition is, however, easily saved by noting that what individuates a computation is not the set of states that the system may occupy, but rather the pattern of transitions between states (Chalmers 1994). Extending this argument from combinatorial state machines discussed by Chalmers to dynamical systems in general, we conclude that intrinsic computation only makes sense if its implementation is extended over *time*.

Another reason to reject the instantaneous-state view of representation and phenomenality in favor of a time-extended one is the inability of the former to articulate the role of units (e.g. neurons) that are not presently active in the present-moment representational or experiential disposition of the system (Fekete & Edelman 2011). To simplify the discussion, and without loss of generality, let us count as active those neurons that are presently undergoing depolarization (i.e. about to emit an action potential). It is hard to see in what sense do the other, presently silent, neurons contribute to whatever it is that the brain is up to at this instant, except counterfactually: were those neurons in fact active, the brain's functional disposition would be different.

The counterfactual stance is, however, not only intuitively unsatisfactory (as noted in the opening section). It also fails to explain what prevents neurons in some other brain, or indeed in the rest of the universe, from counting as part of the system in question. As before, letting time roll solves the problem. In particular, the neurons that were silent a moment ago can now be seen as having contributed to the brain's function factually: *because* they were silent, the system's trajectory through its state space went *this* way, rather than some other way. By the same principle, other brains' neurons (along with the pigeons in Trafalgar Square and the chunks of ice in the rings of Saturn) are excluded, as they should be, because they have not contributed to the shaping of the system's trajectory during the time window in question. This proposal brings us naturally to what would seem to be the next question: how long is this time window?

4. Waiting for Godot

The starting point for taking up this question must be the observation that neurons in a network cannot communicate instantaneously. The transmission delays between the brain's neurons, as determined by the speed of action potential propagation and by the dynamics of synapses and of synaptic potential integration, are non-negligible; they also depend on the neurons' location relative to one another. This basic characteristic of brain networks, which underlies the synfire theory (Abeles 1991) and the related concept of polychronization (Izhikevich 2006),

should by rights be as central to theoretical neuroscience as Einstein's special relativity theory is to electrodynamics. In particular, the unavoidability of time lags in inter-neuron communication implies that any serious theory of brain function must deal with time explicitly (e.g. by treating the brain and its environment as a dynamical system).

Only with time, as the system's far-flung components interact, does its dynamics become apparent – most importantly, to the system itself, that is, intrinsically. Because all the components that can in principle contribute to the system's dynamics participate in shaping its trajectories through the state space and in imposing structure on this space (ruling some classes of trajectories in and others out), phenomenal experience is holistic: it emerges from the dynamics of the entire brain (Fekete & Edelman 2011).⁴ But as we just noted, because signaling within a network of neurons cannot be instantaneous, the holism of experience implies that it too must be inherently temporally extended.

4.1 The attractor hypothesis

Just how should time enter into the would-be dynamical theory of representation and of phenomenal experience? It is tempting to assume that a mere delay between the exposure to a stimulus and its experience, corresponding to a dynamical system settling into an attractor, would suffice. Attractor dynamics plays a key role in the self-observation account of experience, exemplified by the proposal that “phenomenal experience emerges when all relevant neurons in a network are informed about their own population state” (Malach 2007; see also this volume).

The idea that brains engage in self-observation certainly has a great intuitive appeal: is it not true that neurons' first order of business is to communicate information about their local states to their peers?⁵ When offered as an explanation of conscious experience, it also seems to have the advantage of being satisfyingly intrinsic: intuitively, it seems appropriate that the epistemic aspect of conscious phenomenal states – as in my incontrovertible “knowledge” that the tomato in front of me is red – should be rooted in self-observation by the brain.

4. Or even outside it. Insofar as minds are embodied and situated, the dynamical processes that constitute them “spill over” from the brain into the body and to some extent into the rest of the environment (Thompson & Varela 2001; Spivey 2006; Edelman 2008).

5. Somewhat surprisingly, this truism has only recently started to make its way into mainstream theoretical neuroscience, helped along perhaps by the realization of the importance of the brain's “intrinsic functional architecture” (Raichle 2006; Mason, Norton, Van Horn, Wegner, Grafton, and Macrae 2007; Raichle 2010).

The notion of self-observation is not new to theoretical neuroscience: it is central, for instance, to Buszák's (2010) theory of "neural syntax." According to this theory, reader mechanisms – anatomical entities (such as neurons) or functional ones (such as oscillations that synchronize readout) – effectively define momentary cell assemblies (Hebb 1949; Abeles 1991; Ikegaya, Aaron, Cossart, Aronov, Lampl, Ferster & Yuste 2004) upstream from themselves, and thereby implement collective action.

Any attempt to define experience in terms of a *consensus* on the part of some population of neurons must, however, face up to the computational difficulties associated with the problem of attaining global agreement in asynchronous distributed systems. In a nutshell, in realistic distributed systems, of which the brain is but one example, failures such as faulty elements and unreliable communication links conspire to make agreement (e.g. about the value of a global variable) very hard to achieve (Pease, Shostak & Lamport 1980; Lamport, Shostak & Pease 1982).⁶

A further complication is introduced by asynchrony: intuitively, asynchronous communication protocols cannot distinguish failed elements from delayed messages (Kapron, Kempe, King, Saia & Sanwalan 2010). Indeed, with just a single faulty processor, no deterministic asynchronous agreement is possible (Fischer, Lynch & Paterson 1983). Even with randomization, the best known algorithm, which succeeds with probability 1, has an expected running time that is exponential in the number of elements (Ben-Or 1983). Considerations of computational tractability indicate, therefore, that the wait until all the neurons in a network have reached an agreement may be long indeed.

There is, however, an even more fundamental issue with the hypothesis that equates experience with convergent self-observation. On the present account, which we outlined in Section 2, this hypothesis is *conceptually* problematic, because it predicates consciousness on individual neurons' attaining population-level knowledge as the system is falling into an attractor. The dynamics of brain function, including consciousness, is, however, better described by analogy to the performance of an unconducted chamber orchestra (cf. Lloyd's (2011) notion of mind as music). Chamber music would sound terrible indeed

6. In textbooks on distributed computation, this problem is introduced by the way of a story about two Byzantine generals attacking a city from two opposite directions. Knowing that their victory depends on coordination, one of the generals sends a runner to the other one, proposing to attack on the third day at dawn. To be on the safe side, he also requests from the second general a confirmation that the message has been received. The second general complies, but he too requests a confirmation. The city is saved by the infinite regress that ensues.

if it proceeded in fits and starts, with each musician holding a note while waiting to figure out what exactly the others are up to. This, of course, is not how such an orchestra works: the individual performers coordinate with each other dynamically, making the music flow. Likewise, consciousness resides not in what neurons individually know about each other's activities, but rather in what they collectively do.

4.2 The way ahead

The popularity of attractor dynamics with consciousness theorists (e.g. Skarda & Freeman 1987; O'Brien & Opie 1999; Malach 2007) reveals a conceptual shortcoming of the default scenario within which brain dynamics has been traditionally studied, and which our very framing of the time question at the end of Section 3 implicitly relied upon. In this standard scenario, a stimulus is delivered to the subject, who after a certain delay experiences it, while producing or withholding an overt response. This Pavlovian staple of 20th century psychology rests on the assumption that over and above priming, the brain's activity prior to the onset of the stimulus is of little consequence to its "processing." Although it is not often spelled out, this assumption is glaringly obvious in the subtitle found on the cover of the 1961 Science Editions paperback printing of Donald Hebb's *The Organization of Behavior* (1949): "Stimulus and response – and what occurs in the brain in the interval between them."

It is now becoming increasingly clear that this assumption is wrong (Spivey 2006, p. 47).⁷ Given that external events modulate ongoing experience rather than generating it *ab vacuo*, it makes little sense to phrase questions about brain dynamics in terms of the "delay" between the stimulus and its experience, which is, moreover, more profitably viewed not as settling into an attractor and waiting for the next stimulus, but rather as "chaotic itinerancy" through a series of saddle points in the state space (van Leeuwen 2007).⁸

7. Cf. Freeman (2007): "In an awake subject engaged with the environment those driven cortical neurons, even before stimulation, are already self-organized and active through intracortical and extracortical feedback loops."

8. According to van Leeuwen (2007), systems that exhibit chaotic itinerancy "are characterized by fragile attractors corrupted by noise. Regions where such attractors used to be are still preferably visited, as the system is attracted to them. The system will dwell in such regions for some time, and subsequently escape to another part of its state space." A similar view of brain dynamics is offered by Huette and Spivey (this volume).

On the level of system dynamics, therefore, one should inquire about the time scales on which experience unfolds (Rudrauf, Lutz, Cosmelli, Lachaux & le van Quyen 2003, p. 38), so as to determine the characteristic time constants that describe the changes in experience brought about by stimulus onset (e.g. Ito, Nikolaev & van Leeuwen 2007). Crucially, such analysis must be complemented by a study of the structure of the system's trajectory space (Fekete, Pitowsky, Grinvald & Omer 2009; Fekete 2010), which is what determines the structure of the experience that arises from it (Fekete & Edelman 2011).

5. Computational tools for distributed dynamic coordination

The brain's trajectory space is shaped by the kind of distributed dynamic coordination among neurons that we have likened earlier to a music performance by an unconduted orchestra. Possible specific functional reasons for such coordination have been posited in the past (including in the context of consciousness; see Merker, this volume). By what computational means can this coordination be sustained? Unlike the members of a chamber ensemble, who can all hear each other directly, neurons in the brain are on the average many times removed from each other in terms of the number of intervening synapses and signal propagation delay. It turns out that computational tools for mediated dynamic coordination are available, which, moreover, circumvent the tractability problem associated with asynchronous distributed agreement over global variables.

The key computational concept here is temporally extended *diagnosis* of the system's trajectory-space structure by a subset of its elements.⁹ The mathematical basis for this operation is a theorem proved by Takens (1981), according to which a series of measurements of a single variable within a dynamical system becomes with time diagnostic with respect to qualitative structure and quantitative parameters of the system at large (Quyen 2003).¹⁰

9. Cf. Quine (1985): "Mental states, construed as states of nerves, are like diseases. A disease may be diagnosed in the light of observable signs though the guilty germ be still unknown to science."

10. For a cautionary note on the conditions for the applicability of Takens' Theorem, see (Smith 1997).

We note that methods derived from the Takens Theorem receive much attention in empirical studies of complex dynamical systems (including ecosystems, which, like the brain on the relevant time scale, are also spatially distributed; Wilson & Rand 1997). Interestingly, such methods become more effective if multiple “observing” variables are used instead of one (Deyle & Sugihara 2011). These results imply that a clique of neurons can diagnose the dynamics of the rest of the brain, given enough time – just what the task of distributed dynamic coordination requires. Moreover, if the dynamics in question is not too high-dimensional (a plausible assumption, for a number of reasons; Skarda & Freeman 1987; Freeman 2007), the number of samples required for a good approximation is manageable (Deyle & Sugihara 2011). This approach thus absorbs Malach’s (2007) idea of self-observation, without, however, requiring convergence to an attractor, and without equating self-observation with phenomenality – which we identify with the ongoing collective action of the brain’s neurons in shaping its trajectory space (Fekete & Edelman 2011).

6. Summary

In this chapter, we discussed the implications of three constraints on computational theories of phenomenal experience: autonomy, timeliness, and tractability. The three constraints – that phenomenality be intrinsic to the system that experiences it, that its unfolding in time be accounted for, and that it be computable within the distributed architecture of the brain – narrow down the range of plausible contenders for a theory of phenomenal experience. In particular, the hitherto popular approaches rooted in the stimulus-response conception of brain function, which equate phenomenality with self-observation as such and temporality with attractor dynamics, are conceptually and computationally problematic.

In comparison, the sketch of a computational theory of phenomenal experience offered in (Fekete & Edelman 2011) passes the crucial test of being statable exclusively in intrinsic terms, while conforming *prima facie* to the time-related constraints stemming from the mind’s embodiment as a distributed dynamical system. Because experience is massively endogenous and continuous, it must be seen not as convergence to an attractor, but rather as the unfolding of a metastable trajectory through a properly structured space of possible trajectories, as defined by the brain’s dynamics. The self-diagnosis theorems suggest the computational means whereby the distributed dynamic coordination required for this purpose may be sustained.

Acknowledgments

We thank Rafi Malach for many useful and exciting discussions.

References

- Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.
- Baars, B.J. (1988). *A cognitive theory of consciousness*. New York, NY: Cambridge University Press.
- Ben-Or, M. (1983). Another advantage of free choice: Completely asynchronous agreement protocols. In *Proceedings of the Second Annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2730.
- Brown, R. (2006). What is a brain state? *Philosophical Psychology* 19, 729–742.
- Buzsáki, G. (2010). Neural syntax: Cell assemblies, synapsembles, and readers. *Neuron* 68, 362–385.
- Chalmers, D.J. (1994). A computational foundation for the study of cognition. Available online at <http://jamaica.u.arizona.edu/~chalmers/papers/computation.html>.
- Churchland, P.S. (1987). *Neurophilosophy*. Cambridge, MA: MIT Press.
- Churchland, P.S. & Sejnowski, T.J. (1992). *The computational brain*. Cambridge, MA: MIT Press.
- Dennett, D.C. (1991). *Consciousness explained*. Boston, MA: Little, Brown & Company.
- Deyle, E.R. & Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS ONE* 6, e18295.
- Edelman, S. (2008). *Computing the mind: How the mind really works*. New York: Oxford University Press.
- Einstein, A. (1905). Zur Elektrodynamik bewegter Körper. *Annalen der Physik* 17, 891–921. English translation: ‘On the Electrodynamics of Moving Bodies’, in G.B. Jeffery and W. Perrett (1923).
- Fekete, F., Pitowsky, T., Grinvald, A. & Omer, D.B. (2009). Arousal increases the representational capacity of cortical tissue. *Journal of Computational Neuroscience* 27, 211–227.
- Fekete, T. (2010). Representational systems. *Minds and Machines* 20, 69–101.
- Fekete, T. & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition* 20, 807–827.
- Fischer, M.J., Lynch, N.A. & Paterson, M. (1983). Impossibility of distributed consensus with one faulty process. In *Proceedings of the Second ACM SIGACT-SIGMOD Symposium on Principles of Database System (PODS)*, 17.
- Freeman, W.J. (2007). Indirect biological measures of consciousness from field studies of brains as dynamical systems. *Neural Networks* 20, 1021–1031.
- Garey, M.R. & Johnson, D.S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco, CA: W.H. Freeman.
- Hebb, D.O. (1949). *The organization of behavior*. Wiley.
- Ikegaya, Y., Aaron, G., Cossart, R., Aronov, D., Lampl, I., Ferster, D. & Yuste, R. (2004). Synfire chains and cortical songs: Temporal modules of cortical activity. *Science* 304, 559–564.

- Ito, J., Nikolaev, A.R. & van Leeuwen, C. (2007). Dynamics of spontaneous transitions between global brain states. *Human Brain Mapping* 28, 904–913.
- Izhikevich, E.M. (2006). Polychronization: Computation with spikes. *Neural Computation* 18, 245–282.
- James, W. (1890). *The Principles of Psychology*. New York: Holt. Available online at <http://psychclassics.yorku.ca/James/Principles/>.
- Kapron, B.M., Kempe, D., King, V., Saia, J. & Sanwalan, V. (2010). Fast asynchronous Byzantine agreement and leader election with full information. *ACM Transactions on Algorithms* 6.
- Lamport, L., Shostak, R. & Pease, M. (1982). The Byzantine generals problem. *ACM Trans. on Prog. Lang. and Syst.* 4, 382–401.
- Lloyd, D. (2011). Mind as music. *Frontiers in Psychology* 2.
- Malach, R. (2007). The measurement problem in consciousness research. *Behavioral and Brain Sciences* 30, 516–517.
- Mason, M.F., Norton, M.I., Van Horn, J.D., Wegner, D.M., Grafton, S.T. & Macrae, C.N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science* 315, 393–395.
- McDermott, D.V. (2001). *Mind and mechanism*. Cambridge, MA: MIT Press.
- Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences* 30, 63–81.
- Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.
- O'Brien, G. & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* 22, 127–148.
- Pease, M., Shostak, R. & Lamport, L. (1980). Reaching agreement in the presence of faults. *J. ACM* 27, 228–234.
- Quine, W.V. (1985). States of mind. *The Journal of Philosophy* 82, 5–8.
- Quyen, M.L.V. (2003). Disentangling the dynamic core: A research program for a neurodynamics at the large-scale. *Biol. Res. [online]* 36(1), 67–88.
- Raichle, M.E. (2006). The brain's dark energy. *Science* 314, 1249–1250.
- Raichle, M.E. (2010). Two views of brain function. *Trends in Cognitive Sciences* 14, 64–71.
- Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-F. & le van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco Varela's exploration of the biophysics of being. *Biological Research* 36, 27–65.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417–424.
- Shanahan, M. (2010). *Embodiment and the Inner Life*. New York, NY: Oxford University Press.
- Skarda, C. & Freeman, W.J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* 10, 161–195.
- Sloman, A. & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies* 10, 113–172.
- Smart, J.J.C. (2004). The identity theory of mind. In E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Stanford University. Available online at <http://plato.stanford.edu/archives/fall2004/entries/mind-identity/>.
- Smith, L. (1997). The maintenance of uncertainty. In *Proceedings of the International School of Physics "Enrico Fermi"*, Volume Course CXXXIII, Berlin, pp. 177–246. Springer.
- Spivey, M.J. (2006). *The continuity of mind*. New York: Oxford University Press.

- Takens, F. (1981). Detecting strange attractors in turbulence. In D. Rand and L.-S. Young (Eds.), *Dynamical systems and turbulence*, Lecture Notes in Mathematics, pp. 366–381. Berlin: Springer.
- Thompson, E. & Varela, F. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences* 5, 418–425.
- Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242.
- van Leeuwen, C. (2007). What needs to emerge to make you conscious? *Journal of Consciousness Studies* 14, 115–136.
- Wilson, H.B. & Rand, D.A. (1997). Reconstructing the dynamics of unobserved variables in spatially extended systems. *Proceedings of the Royal Society of London B* 264, 625–630.

The (lack of) mental life of some machines

Tomer Fekete & Shimon Edelman

Department of Biomedical Engineering, Stony Brook University / Department of Psychology, Cornell University

The proponents of machine consciousness predicate the mental life of a machine, if any, exclusively on its formal, organizational structure, rather than on its physical composition. Given that matter is organized on a range of levels in time and space, this generic stance must be further constrained by a principled choice of levels on which the posited structure is supposed to reside. Indeed, not only must the formal structure fit well the physical system that realizes it, but it must do so in a manner that is determined by the system itself, simply because the mental life of a machine cannot be up to an external observer. To illustrate just how tall this order is, we carefully analyze the scenario in which a digital computer simulates a network of neurons. We show that the formal correspondence between the two systems thereby established is at best partial, and, furthermore, that it is fundamentally incapable of realizing both some of the essential properties of actual neuronal systems and some of the fundamental properties of experience. Our analysis suggests that, if machine consciousness is at all possible, conscious experience can only be instantiated in a class of machines that are entirely different from digital computers, namely, time-continuous, open, analog, dynamical systems.

1. Introduction – special laws

The hypothetical possibility of building a sentient machine has long been a polarizing notion in the philosophy and science of mind. The computer revolution and the emergence in the last decade of the 20th century of scientific approaches to studying consciousness have sparked a renewed interest in this notion. In this chapter, we examine the possibility of machine consciousness in light of the accumulating results of these research efforts.

Under a liberal enough definition, any physical system, including a human being, can be construed as a machine, or, indeed, a computer (Shagrir 2006). Moreover, the concept of consciousness itself turned out to be very broad, ranging from minimal phenomenal awareness or sentience (Merkur 2007) on the one extreme to higher-order thought on the other (Rosenthal 2005). We shall, therefore, focus our analysis on a narrow, yet fundamental, version of the machine

consciousness question: whether or not digital computers can have phenomenal experience of the kind that one intuitively attributes to any animal with a brain that supports sensorimotor function. To that end, we restrict our consideration to digital *simulations* of brains, construed for present purposes simply as networks of biological neurons.

To anticipate the thrust of our inquiry, if it turns out that a digital computer that simulates a brain is categorically precluded from having a phenomenal life, the idea of computer consciousness would be effectively doomed, given that brains are the only example we have of conscious “machines”. If, on the contrary, one manages to show that a digital simulation of a conscious brain suffices to give rise to consciousness in its own right, this would amount to the discovery of a unique kind of natural law. To date, all natural laws discovered through scientific endeavor are stated in terms of mathematical equations that relate *physical* properties of matter (e.g. mass, electrical charge, etc.). In contrast, a digital simulation is an instantiation of an algorithm, and as such is by definition multiply realizable, that is, it depends not on the physical composition of the system that implements it but rather on its formal organization.

The principle that underlies the alleged ability of computer-simulated brains to give rise to experience, as stated by Chalmers (1995), is **organizational invariance (OI)**, according to which “*experience is invariant across systems with the same fine-grained functional organization. This is best understood as the abstract pattern of causal interaction between the components of a system, and perhaps between these components and external inputs and outputs. A functional organization is determined by specifying (1) a number of abstract components, (2) for each component, a number of different possible states, and (3) a system of dependency relations, specifying how the states of each component depends on the previous states of all components and on inputs to the system, and how outputs from the system depend on previous component states. Beyond specifying their number and their dependency relations, the nature of the components and the states is left unspecified. ... I focus on a level of organization fine enough to determine the behavioral capacities and dispositions of a cognitive system. This is the role of the “fine enough grain” clause in the statement of the organizational invariance principle; the level of organization relevant to the application of the principle is one fine enough to determine a system’s behavioral dispositions.*”

Properties and phenomena that exhibit OI differ from those governed by natural laws that are familiar to us from physics. The merit of a physical theory lies in the goodness of fit between the formal statement of the theory – a set of equations – and the relations among various physical measures. As such, a theoretical account of a physical phenomenon is always associated with a degree of approximation, the reasons for which may be technological (the accuracy of measurement devices),

numerical (stemming from the mathematical approach adopted by the theory), or fundamental limitation (as in quantum indeterminacy).

The situation is markedly different with regard to OI properties, because it is a priori unclear how the implementational leeway allowed by OI should relate to explanatory accuracy. Thus, a model of a brain can be very successful at describing and predicting various physiological measures (e.g. membrane potentials), yet still be open to doubting whether or not it captures the fundamental formal properties of the brain that supposedly realize consciousness. The stakes are particularly high in any attempt to formulate an OI explanation for conscious experience, where the explanandum is definite and specific in the strongest possible sense: experience, after all, is the ultimate “this.”

Indeed, it is not enough that a digital simulation *approximate* the function of the sentient neural network: it must capture all the properties and aspects of the network’s structure and function that pertain to its experience, and it must do so precisely and in an intrinsic manner that leaves nothing to external interpretation. In other words, the network that is being modeled, along with its ongoing experience, must be the unique, intrinsic, and most fundamental description of the structure of the simulation system (comprising the digital simulator and the program that it is running), and of its function. If that description is not the most fundamental one, nothing would make it preferable over alternative descriptions. If it is not unique, the simulator system would seem to be having multiple experiences at the same time. If it is not intrinsic, the simulator’s experience would be merely an attribution. Any of those failures would invalidate the claim that the simulator does the right thing with regard to that which it purports to simulate.

In what follows, we analyze neural network simulation in light of these challenges, which raise the ante with regard to the possibility of a digital computer emulating an actual brain *qua* the substrate of the mind. We begin by analyzing the neural replacement scenario, which has been the source of some of the most powerful arguments in favor of the possibility of digital minds.

2. One bit at a time

In neural replacement scenarios, one is asked to imagine that a brain is replaced, typically one neuron at a time, by a digital functional equivalent (e.g. Pylyshyn 1980). The extreme case, in which each of the brain’s neurons is replaced by a full-fledged digital equivalent, amounts to simulating brain in a digital system. Note that if we omit “digital” from this description, we are left with a statement that is neither controversial, nor, alas, informative: it simply begs the definition of functional equivalence. The crucial issue here is the very possibility of a functionally

equivalent surrogate, be it digital (e.g. an electronic circuit), biological (e.g. a neuron resulting from manipulating stem cells) or what have you.

This issue is sidestepped by Chalmers (1995), who writes, “*We can imagine, for instance, replacing a certain number of my neurons by silicon chips. ... a single neuron’s replacement is a silicon chip that performs precisely the same local function as the neuron. We can imagine that it is equipped with tiny transducers that take in electrical signals and chemical ions and transforms these into a digital signal upon which the chip computes, with the result converted into the appropriate electrical and chemical outputs. As long as the chip has the right input/output function, the replacement will make no difference to the functional organization of the system.*”

The preceding passage equates the local function of a neuron with the input/output function of a digital chip. In terms of the abstraction step that is part and parcel of any claim of OI, it abstracts away every aspect of the replacement or the simulation except its input/output relations. Note that the digital replacement/simulation (DR) scenario fixes the level of description (resolution) of the functional specification of the replaced/simulated (sub)system at some definite spatiotemporal level: if the chip itself is described according to the guidelines above – that is, by enumerating the digital chip’s parts, possible states, etc., according to the prescription of OI – it rules out the possibility of *complete* functional identity to a neuron.

Setting aside the question of whether or not the neuronal level itself is the fundamental level for understanding the brain, we would like to explore the consequences of making this theoretical move, that is, setting a definite *categorical* threshold for functional equivalence (an alternative would be to conceive of functional correspondence as graded, with the *degree of similarity* replacing all-or-none equivalence). To put it bluntly, could it really be the case that a description up to a certain level of organization is a fundamental constituent of reality, yet beyond that point all details are inconsequential?

With these observations in mind, let us analyze the DR scenario carefully. To begin with, it should be noted that if DR preserves functionality, then the inverse process – going from a full-fledged simulation to a real brain one neuron at a time – must do so too. This is important, because even if one carries out a physiologically informed analysis, the actual burden each neuron carries might escape notice if only the first step of DR – replacing a single neuron – is considered. Simply put, surrogate neurons (SNs) must function in a way that would keep actual neurons’ function unchanged even if these are an overwhelming minority within the DR.

A surrogate neuron must share the approximate morphology of the original cell. The SN must extend spatially through all the synapses – incoming and outgoing alike – though which the original neuron was joined to the rest of the

network. Neurons are not point switchboards; rather, the functional role of the connection between two neurons depends on the exact location of the synapse onto the post-synaptic cell. Not only does the impedance of the neuron's parts vary greatly as a function of local cell morphology, but the signal propagation (synaptic) delay too varies considerably as a function of distance from the cell body. Moreover, each axon synapses at multiple points on different cells, and similarly dendritic trees receive thousands of synapses. This means that the entire dendritic tree and axonal ramifications of the SN would have to match exactly those of the original neuron, otherwise both the timing and strength of the inputs would be profoundly altered.

All the presynaptic terminals of an SN must have some DNA-based metabolic functionality. The presynaptic terminals do not only secrete neurotransmitters and neuromodulators into the synaptic clefts: a crucial part of their function is neurotransmitter reuptake. Without reuptake, neurotransmitters remain active (that is, they can bind to postsynaptic receptors) for extended periods (this is the principle behind the action of SSRI, or selective serotonin reuptake inhibitors, as antidepressants – altering the kinetics of reuptake results in profound change in function). In other words, reuptake is part of what establishes the actual “output” of a synapse. Moreover, neurotransmitters have a limited functional life span before becoming corrupted by various chemical and metabolic processes. Hence, neurotransmitters must be constantly broken down, and new proteins and peptides synthesized in their stead.

Thus for the SN to fulfill its intended function, either cell-like entities would need to be maintained at each terminal, or the SN would need to take the form of a single, spatially extended cell entity, which, as noted above, would share the morphology of the original cell. The former option would imply that what Chalmers nonchalantly refers to as digital/analog transduction would thus remain an unsolved problem – these little metabolic machines that were supposed to act as transducers would still have to be interfaced with. It is not at all clear how this could be achieved, but it would have to involve the digital signal controlling calcium influx into the cell (which is part of the natural cascade of events leading to transmitter release). Seeing that this is not the only function these cell-like entities would need to carry out, this option seems even less defensible.

SNs would have to consist at least in part of closed compartments that would maintain the normal electrochemical gradient across the membranes of the brain's cells. Neurons achieve their functionality by actively maintaining a marked difference in ion concentration – and thus electrical potential – between their insides and the extracellular space. This is done through pumps and channels – pipe-like proteins, which through their structure, and at times through energy expenditure, control the ionic flow across the membrane (e.g. negative in, positive

out). An action potential or spike results from opening the “floodgate” and allowing ions from the extracellular space to rush in. This event self-terminates through the locking of some voltage-dependent channels, as well as through the equalization of the cross-membrane potential, and is accompanied by a concentrated effort to expel positive ions to restore the cell’s resting potential.

Some neurons are electrically coupled through gap junctions on their dendrites, somas or axons (Connors & Long 2004; Wang et al. 2010). These are channel-like structures that allow bidirectional flow of ions and at times proteins and other molecules between two neurons. Gap junctions are thought to be responsible for the synchronized firing capacity of neuronal circuits, and possibly for the formation of cell assemblies – both properties that are believed by many to be essential to neuronal representation (Milner 1974; Hebb 1988; Von Der Malsburg 1994). Thus, a SN would need to maintain compartments employing pump- and channel-like mechanisms to mimic the ion dynamics of gap junctions.

It might be argued that, as gap junctions are very sparse in certain cortical areas among primary neurons (although not interneurons), such compartments would be few and far in between, and thus perhaps could be treated as digital to analog transducers. However, one cannot dismiss out of hand the possibility that SNs would need to mimic the electrochemical dynamics in full – as otherwise the ionic distribution in the extracellular space at large might be profoundly altered, rendering real neurons ineffectual, once more than a handful of neurons are replaced.

Without the assumption that glial cells are merely “housekeeping” elements, the DR scenario seems considerably less plausible. Glial cells are massively interconnected through gap junctions (Bennett & Zukin 2004), and, moreover, not only engage in various signal exchanges with neurons but are in fact coupled to neurons through gap junctions as well (Alvarez-Maubecin et al. 2000). In computational terms, glia at the very least establish the connectivity parameters of the network formed by the brain: even if the intricate calcium dynamics generated by glial networks are somehow shown to have no role in cognition (for contrary evidence see, e.g. Scemes & Giaume 2006), it is still true that glia are responsible to a large extent for maintaining the functionality of neurons, by affecting transmission delays through myelination, and by contributing to the control of synaptic efficacy (connection strength; (Fields & Stevens 2000; Shigetomi et al. 2008; Theodosis et al. 2008; Ricci et al. 2009; Eroglu & Barres 2010; Pannasch et al. 2011). If all this is mere housekeeping, so is the semiconductivity of doped silicon in digital chips.

Thus, if glia are taken to be an essential part of the functional network that DR is to emulate, an SN would have to comprise numerous compartments in which the original electrochemical dynamics are mimicked, to the point it might

be necessary to recreate a single compartment shaped approximately as the original cell and implementing the full original electrochemical dynamics, so as to maintain functional invariance both at the level of cells and at the network level.

DR is particularly problematic in developing brains. In the process of maturation, gap junctions are more prevalent and seem to play an essential part in producing patterns of spontaneous activity necessary for the formation of the connectivity matrix between neurons. Moreover, at this stage there are massive rewiring processes going on. This situation makes it even more important that SN replicate the electrochemical dynamics and membrane functionality of the original neurons.

In summary, a putative SN would have to be a far cry from a digital circuit equipped with a few transducers. From known physiology, one can make a strong case that a successful surrogate would have to approximate the morphology of the original neuron, engage in DNA-based synthesis and regulation of many metabolic processes, and maintain an intracellular environment and electrochemical dynamics nearly identical to the original. The DR scenario could thus actually be used as an argument in support of “neural chauvinism” (Block 1980). Even if we accept this scenario, a more appropriate name for a conglomerate of cell-like entities, shaped to resemble neurons and to function as they do, would be a cyborg or a chimera, rather than a silicone chip, even if the entire contraption is in some limited sense controlled by one. Just as importantly, it would then be hard to see in what sense the computations carried out by the digital component of the SN are more abstract or fundamental compared to what the rest of the SN contraption does. If the original electrochemical and biochemical dynamics would have to be replicated by the SN, it makes sense to argue that these processes, rather than the digital computation, are the fundamental realizers of consciousness. Thus, the entire thought experiment of DR is rendered inconclusive at best.

The applicability of the idea of organizational invariance to brain-based cognition and consciousness appears therefore to be suspect in two respects. First, the notion of abstraction that is fundamental to OI, which focuses on input-output relations, seems inappropriate, or at least too general to be of theoretical or practical use. Second, OI is revealed to be fundamentally inconsistent when considered in detail. On the one hand, it assumes that minds share with computations (but not with any kind of physically defined entity or process, for which the implementation matters) the key property of being definable via a description alone. On the other hand, rather than offering measurable criteria for goodness of fit between descriptions (models) and their instantiations, OI attempts to marginalize this critically important issue (e.g. by denying the relevance of neural details that seem problematic). This double-standard stance casts a doubt on the entire notion of there being two classes of phenomena – one that is uniquely determined

by the regular laws of physics and the other, including consciousness, subject to a unique kind of natural law in which only some arbitrary level of description matters – which is central to the OI-based theories of consciousness.

3. How detailed is detailed enough?

Given the difficulties inherent in the attempts to make organizational invariance, or OI, work via the digital replacement scenario, we would like to consider next the purely digital scenario – namely, simulating an entire brain, in as much detail as needed – and see if OI fares better there. As always, we have two questions in mind: (1) does a simulation in fact instantiate the functional architecture of the brain (neural network) that it aims to simulate, in the sense that this network model indeed constitutes the best description of the underlying dynamics of the digital machine, and (2) what are the implications of deciding that the simulation needs only to be carried out down to some definite level, below which further details do not matter?

Brain simulation is not an isomorphism but rather a partial intermittent fit. To satisfy the conditions of OI, the instantiation of an algorithm realizing brain dynamics equations must be *isomorphic* to the causal organization of the brain at some fundamental level of functional, spatial, and temporal resolution. By definition, an isomorphism is a bijective mapping between two domains that preserves essential structure, such as the computational operations that sustain the causal interactions.

In the present case, a digital computer simulation is most definitely not isomorphic to brain dynamics, even if correspondences only above a certain level of resolution are considered. Indeed, the mapping, to the extent that it holds, only holds at certain points in time, namely, whenever the machine completes a discrete step of computing the dynamics. In the interim, the correspondence between the dynamics of the simulation and of its target is grossly violated (e.g. by duplicating various variables, creating intermediate data structures, and so on).

Furthermore, due to the finite precision of digital computers, all such computation is approximate: functions are approximated by the leading terms in their Taylor expansion, and complex system dynamics are approximated numerically using finite time increments and finite-element methods. Thus, not only is correspondence intermittent, but it only applies to inputs (up to the round-off error of the machine); simulated outputs are at best ϵ -similar to actual outputs (especially as the time progresses from one “check point” to the next). In fact, if the simulated system of equations cannot be solved analytically, we cannot even gauge to what extent the simulacrum and the real thing diverge across time.

Moreover, our limited capacity to measure brain dynamics densely in space and time, coupled with a limited understanding of what we do manage to measure, holds us back from corroborating the extent to which a simulation is sufficiently similar to the original.

Unfortunately, the OI idea itself does not offer any means for assuaging these concerns. Clearly, if the simulation is not precise enough, it will not be up to the task – after all, any arbitrary system can be construed (in a perfectly counterfactual manner) as a bad simulation of a brain. What then constitutes an adequate simulation? Even for the quantities that a digital simulation has some hope of getting right (namely, inputs and states that occur at some chosen points in time), the correspondence is only partial (for example, the bits in the total state of the machine that are responsible for instantiating the operating system are assumed tacitly to be immaterial to its mental life). One thus begins to suspect that simulation, rather than being an organizational invariant, is a reasonable and robust mapping between the simulated system and the simulating one, whose detailed operation may be far from inconsequential. Alas, the mapping between the text on this page and the syntactical structures it encodes is another such mapping, yet its representational properties are nil in the absence of an external reader/interpreter – a problem for which brain simulation was supposed to be a solution rather than an example.

Our abstractions are not as abstract as we think. Imagine a computer running a 3D design program, displaying on the screen the design for a coffee mug. Now, imagine further that rather than having a modern LCD display, the display in this case comprises a square array of photodiodes capable of emitting light in the entire visible spectrum. If on a whim one were to rearrange these diodes randomly, the image on the screen would no longer look like a rotating cup. Clearly, the computation instantiated by the computer in both cases would be one and the same. It could be argued that this is immaterial, as what is fundamental are the inner relations and transformation carried out on the variable arrays – representing a cup in this case – and that the display is just an aid to make all this visible to us.

We contend however that this scenario points at something fundamental not only to visual representations (e.g. drawings and graphical displays), but to symbolic representation (and the associated thought processes) in general. Specifically, people tend to downplay the extent to which symbolic representations (e.g. equations) actually need to undergo an additional interpretative process for the correspondence between the instantiating physical tokens (e.g. writing, or numeric arrays in a computer) and processes (transformation of representations – e.g. addition) and the formal structure they purport to realize to hold. Are brain simulations any different?

Let us look carefully at what happens in a simulation. The initial step is casting the system of equations that describes the process that is to be simulated into a language-like form – usually in a high-level programming language. This transforms the abstract symbolic equations into an algorithm – a set of procedures that corresponds to the equations in the sense that their inputs and outputs can be mapped bijectively to the symbols in the equations. The algorithm, in turn, needs to be translated into executable code, in a process of compilation that usually involves multiple steps (e.g. Matlab to Java to assembly language to machine code).

In the above example we saw that arranging the outputs of a computation spatially with appropriate color coding allows us to further interpret a computation, that is, impose additional structure – e.g. a rotating cup – on an instantiating process by putting our perceptual interface into play, a process that is certainly not unique. Why is it then that the translation from machine codes to various programs, then to an algorithm, and finally to equations (which need an observer to be interpreted) is expected to be unique and intrinsically determined, unlike the translation of a pixel display into the visual concept of a cup, which is far from intrinsic or unique? To see that this is indeed a tall order that digital simulation cannot meet, let us consider some of the details of representation of numbers in digital systems.

Digital computers usually represent numbers in a binary format, whose details, however, may vary from system to system. One of the choices that must be made is how to represent the sign of the (necessarily finite-precision) number. The options developed to date are one’s complement, two’s complement, signed magnitude, and so on. Table 1 summarizes these options, as well as the basic unsigned interpretation option, in the case of 8-bit representations:

Binary value	Ones’ complement interpretation	Two’s complement interpretation	Signed interpretation	Unsigned interpretation
00000000	+0	0	+0	0
00000001	1	1	1	1
...	...	...	...	...
01111110	126	126	126	126
01111111	127	127	127	127
10000000	−127	−128	−0	128
10000001	−126	−127	−1	129
10000010	−125	−126	−2	130
...	...	...	...	...
11111110	−1	−2	−126	254
11111111	−0	−1	−127	255

Let us look at how a system that uses one's complement architecture carries out the operation $127 + (-127) = 0$. What actually happens is that two binary numbers are mapped to a third one: $(10000000, 01111111) \rightarrow 00000000$. The very same mapping under the two's complement rules means $127 \circ (-128) = 0$; under the signed magnitude rules $127 \circ (-0) = 0$; and under the "vanilla" binary rules $127 \circ 128 = 0$, where $\circ$ in each case stands for some "mystery" binary mapping. While it is computable, $\circ$ looks very much unlike addition under all those interpretations.

This example demonstrates that higher level organization of binary operations – as would be required, e.g. for a simulation of the brain – is highly contingent on interpretation. Under one interpretation, a simulation may seem to realize the neuronal dynamics it was programmed to. At the same time, under other interpretations, it would realize a slew of rather convoluted series of logical and numerical operations, which certainly do not self-organize on a coarser grain into a semblance of a neural net. If so, why is it reasonable to assume that the higher level description of the simulation is somehow inherent to it, while admitting that the cup in the above example is in the eye of the beholder? Could one seriously try to argue that the program can be reconstructed from the digital dynamics with no additional information? Why is it then that one interpretation is deemed to be inherent to the machine dynamics and, indeed, fundamental, while another – say, the "straight up" total binary machine dynamics, which satisfies the same formal constraints and then some – is not? Because the first one makes more sense to us? Because the manual provided by the hardware manufacturer says it is the right one?

Perhaps the interpretation recommended by the hardware manufacturer happens to be the simplest, most parsimonious description of the computer's dynamics, and is therefore deserving of the title "real"? Alas, this is not the case. Let us look for example at the two's complement architecture (which is the commonly used one at present). The key rule for interpretation under this format is this: given 2^n representing tokens, all those smaller than 2^{n-1} (the radix) are taken to be "ordinary binary numbers"; the rest are taken to be negative numbers whose magnitude is given by flipping the bits, and adding a least significant bit (see table). Unlike under the one's complement format, there is no distinction here between $+0$ and -0 , which implies that the sign flip operation is defined differently for 0 compared to other numbers. There is nothing particularly parsimonious or natural about this convention.

The problem of the causal nexus. We stated above that simulation is in fact a partial intermittent fit. Another way of phrasing this is to note that one of the fundamental differences between the simulated and the simulation is the way the casual interactions are instantiated. A dynamical system is defined through enumerating components, their possible states and the pattern of causal interactions between

these elements. While the state of the elements is actual (it is measurable), the pattern of causal interactions can only be assessed indirectly through various relations between the pertinent variables. However, in simulation the situation is markedly different; the dynamical equations – a formal representation of the pattern of causal interactions – are explicitly encoded alongside the various parts (variables) the system comprises. While Chalmers (1995) tries to circumvent this issue by introducing the notion of grain, the fact of the matter is that causal interactions do in part fall under the same description as the simulated dynamics – namely, they are realized through components with several possible states at the same grain of the simulation (i.e. representations of numbers as arrays of bistable memory cells), and not simply by operations (flipping bits, local gating action, writing into and reading from memory). Let us see what price is exacted by this somewhat inconsistent policy of machine state interpretation.

The causal nexus can have profound effect on the formal properties of the realized dynamics. A digital simulation of a neural network proceeds by computing periodic updates to the state of the network, given its assumed dynamics (equations), the present state, and inputs. This is achieved by incrementing each variable in turn by the proper quantity at every time step Δt . If the state of the network at time t is described by a vector $x(t)$, then the updates will have the form $x_i(t + \Delta t) = x_i(t) + \Delta x_i(t)$, where $\Delta x_i(t)$ is obtained by solving the system equations. If implementation details are immaterial as long as high-level algorithmic structure remains invariant, all orderings of the updates would be equivalent, including a random one. This observation has implications for the structure of the state space of the simulated dynamical system.

For simplicity, let us consider a system whose dynamics is confined to a low-dimensional manifold. Assume, for instance, that the set of all possible states of the dynamical system $\{x(t)\}$ is topologically a ring. However, if we look at the set of all possible states of our instantiation, we find that instead of a point (i.e. the state $x(t_i)$) we actually have a manifold of nontrivial topological structure. To see this, consider two “adjacent” points on the ring, which without loss of generality we will assume to be 0 and $\Delta x \in \mathbb{R}^3$. Thus, if we consider the set of all possible states of the instantiation, where we once had only the points 0 and Δx , we now have the set including in addition all possible intermediate states under random update, i.e.

$$\Gamma = \left\{ \begin{bmatrix} 0 & \Delta x_1 & 0 & 0 & \Delta x_1 & \Delta x_1 & 0 & \Delta x_1 \\ 0, & 0, & \Delta x_2, & 0, & \Delta x_2, & 0, & \Delta x_2, & \Delta x_2 \\ 0 & 0 & 0 & \Delta x_3 & 0 & \Delta x_3 & \Delta x_3 & \Delta x_3 \end{bmatrix} \right\},$$

where Δx_i are the coordinate updates according to the dynamics. This collection of points is a discrete sample of the surface of a cube, which is not topologically

simple. This is akin to starting off with a string necklace (the topological ring of possible states) only to discover that it magically transformed into a (hollow) bead necklace.

Continuing in a similar fashion would result in further elaboration to the structure of the possible state space. Rather than incrementing the dynamics by Δx_i , we can make the system undergo coordinates increments of $k\Delta x_i/m, k=1, \dots, m$. While this would be somewhat silly computationally, we would nevertheless be meeting the criteria laid down by OI – matching of inputs and outputs (every m steps of the dynamics). In this case, instead of each point in the original ring, we would now have a discrete sample of the edges of a cube, which are topologically distinct from the surface of a cube resulting from the above example. In this scenario our beads would be elegant wireframe beads.

If one concedes that these initial steps somewhat mar the goodness of fit between the original system equations and the actual implementation, he would have to bite the bullet by conceding further that if so, then obviously the straight-up no-funny-stuff original instantiation is better described by asynchronous dynamical equations, i.e.

$$x_i(t) = \begin{cases} x_i(t) + \Delta x_i & \neg \text{mod}(t - i\Delta t, N) \\ x_i(t) & \text{mod}(t - i\Delta t, N) \end{cases}$$

where as before Δx_i are the coordinate updates according to the “real” dynamics and N the total number of state variables.

The easy way out of course, would be to argue that the structure of the state space is immaterial to realizing consciousness. However, that cannot be the case. To see this, one must recall that under OI, the set of all possible states is exactly the set of all possible experiences. This means that the structure of the state space reflects the structure of the conceptual/perceptual domain realized by the simulated brain. Thus the similarity between points embodies the relations between percepts and mental content at large. Therefore a marked difference in the topology of the state space would indicate that in fact the conceptual domain realized by the simulation would have a more elaborate structure than the original.¹

One could try and defuse this problem by appealing to a metric notion, under which these differences would exist but would be negligible relative to other prominent structural facets of the simulations state space. Alas, OI, apart from being

1. In fact, things are actually much worse, as in this example we highlighted the “cleanest” of the array variables realized as part of the “causal nexus” – namely those which are almost identical to the simulated array, while overlooking the slew of various other arrays necessary for interim computations.

fundamentally non-metric, would again seem to claim that some formal properties are fundamental constituents of reality, while what happens below some arbitrary level of consideration has no significance whatsoever. Unfortunately, our analysis seems to indicate that the only pertinent formal properties of the simulation are a partial fit of some of the machine states, mostly an input-output matching affair (or rather an input-input matching affair).

Under OI, at least some brain simulations would realize more than one mind. While simulating the dynamics of a system, regardless of algorithm, the array of state variables (e.g. membrane potentials) has to be duplicated, in order to enable update $x(t + \Delta t) = f(x(t))$. Let us look at time t_{i+1} at which the computation of $x(t_{i+1})$ is complete. Given that OI admits partial intermittent mapping between computer and brain states, in the series of partial states that obtain at times $\{t_1, t_2, \dots, t_p, t_{i+1}\}$ there must be a partial machine state that corresponds to $x(t_{i+1})$ and another one that corresponds to $x(t_i)$. Moreover, still within the OI framework, we observe that the pattern of causal interactions between the elements of x is exactly that of our simulated brain, regardless of what moment of time is used as a reference to describe the series. Thus, if one of the two time series gives rise to a mind, so must the other, which implies that the same simulation instantiates two (slightly time-lagged) minds, none of which is in any intrinsic way privileged over the other.

The slipperiest of slopes. If the principle of organizational invariance applies to simulation of brain function, it must apply also to the brain itself. Clearly, my brain is functionally equivalent to my brain minus a single neuron: taking out a single neuron will not change my behavioral tendencies, which, according to Chalmers (Chalmers 1995), amounts to OI. This is clearly evidenced by neuronal death, a commonplace occurrence that goes unnoticed. Suppose there is, as per OI, a critical level of description of a brain, which we may assume without loss of generality to be the neuronal level. A disturbing consequence of this assumption is that each brain would seem to instantiate at least N (admittedly virtually identical) minds, N being the number of its neurons (or their parts times their number, if a sub-neuronal level is posited to be the critical one).

This, however, cannot be the case. The reason is that under the assumption that we just made, your mind is one among the multitude realized by the brain composed of N neurons. Tragically, when a single neuron out of this bunch dies, the existence of the minds sharing that neuron terminates at that instant. Thus, the fact that you, the reader, are reading this is nothing short of a miracle. Of course, if the functionality of the brain is construed as belonging in a continuous dynamical system space, no such problem arises. In that case, similarity between brains and brain-like systems can be defined in a graded fashion, and can therefore accommodate both growth and degeneration (Fekete 2010).

4. Not all machines are born equal

The previous section leveled many arguments against the notion that digital simulation actually instantiates the neuronal network model supposedly carrying the brunt of realizing consciousness. However, we will need a much humbler proposition to move forward – namely, that a description of the dynamics carried out by a computer simulating a brain in terms of a finite binary combinatorial automaton is at least on par with other formal schemes, such as the neuronal network model, purporting to do the same. If so, we would like to directly compare some of the properties of the total machine state dynamics during digital computer simulation, first with the dynamics of actual brains, and then with the dynamics (or rather ebb and flow) of phenomenal experience.

To that end, we introduce the notion of *representational capacity* (Fekete 2010; Fekete & Edelman 2011). A dynamical system gives rise to an *activity space* – the space of all possible spatiotemporal patterns a system can produce. Such spatiotemporal patterns can be conceptualized as trajectories through the system's (instantaneous) state space. A fundamental constraint on the organization of the activity trajectory space of an experiential system is suitability for capturing conceptual structure: insofar as phenomenal content reflects concepts, the underlying activity must do so as well. The basic means of realizing conceptual structure is *clustering* of activity: a representational system embodies concepts by parcelling the world (or rather experience) into categories through the discernments or distinctions that it induces over the world.² As it gives rise to experience, qua instantiating phenomenal content, activity should possess no more and no less detail than that found in the corresponding experience. Specifically, activities realizing different instances of the same concept class must share a family resemblance (Wittgenstein 1953), while being distinct from activities realizing different concepts. This means that the activity space must divide itself intrinsically into compartments, structured by the requisite within- and between-concept similarity relations.

Furthermore, the richness of experience varies greatly not only between species, but can in fact vary due to change in state of consciousness or *experiential state*; from full-fledged richness in alertness, through dimness (e.g. on the verge of sleep), or be entirely absent (e.g. dreamless sleep, anesthesia). Note that the notion of experiential state pertains to neural simulation as well, that is, if a

2. In terms of experience, distinctions made at the operational level are manifested as differentiation in the phenomenal field (everything that makes up awareness at a given moment). If, say, two different odorants evoke indistinguishable percepts, the underlying activities must have been indistinguishable (in the metric sense) as well.

neural simulation indeed gives rise to experience, this would apply to its activity as well – in this case experiential state would be realized by (and hence correspond to) change in various parameters of the simulation (e.g. those corresponding to levels of certain neuromodulators).

The crucial point here is that the richness of the experience realized by a system corresponds to the degree to which its activity separates itself into clusters. The reason is simple: the more clustered the system's activity, the more distinctions it can draw. Moreover, activity being *the* realization of experience, it is not supposed to require any further interpretation. In other words, activity must impose structure on experience intrinsically, or not at all. Accordingly, if a system does not exhibit intrinsically clustered activity, it cannot be engaging in the representation of its environment in any interesting way, as its activity does not in itself induce any distinctions, and hence its phenomenal field (i.e. everything that makes up its awareness at a given moment) remains undifferentiated. Consider a system that gives rise to a homogeneous activity space: say, its activity is equally likely to occupy any point inside an n -dimensional cube (n being the number of degrees of representational freedom of the system). Such a homogeneous volume in itself does not suggest any partitioning, and any division of it into compartments would be arbitrary. Thus, the activity of this system cannot amount to experience.

Various subtler distinctions concerning the structure of clusters can be made and quantified. One important issue here is the hierarchical structure of clusters (clusters of clusters and so on). In the case of conceptual structure, hierarchy is a means of realizing dominance or inclusion relations among concepts. Other important relations can be modeled by the spatial layout of clusters in the activity space. For example, visual objects can be distinguished according to several parameters such as shape, color, texture, etc., which may be represented by various dimensions of the activity space. Similarly, subdomains of conceptual structures may vary in their dimensionality. Accordingly, the local *effective dimensionality* of configurations of clusters in the activity space is crucial in realizing a conceptual domain.

If so, what are the systematic structural changes in activity that correspond to, say, going from dreamless sleep all the way to full wakefulness? If systematic change in the richness of experience corresponds to a change in experiential state, the richness of experience remains constant when the experiential state is fixed. We can say then that given an experiential state, the complexity of experience is invariant, and so must be the complexity of activity trajectories. What happens when the experiential state changes?

As one emerges from the oblivion of dreamless sleep, one is able to take in more and more details of the surroundings. To do so, the system must be able to make finer and finer discernments regarding both the internal and external

environment. A change in experiential state is thus associated with change in the conceptual structure realized by activity trajectories. At the same time, as experience becomes richer, and with it the realized conceptual domain, the structure of activity trajectory space, which encompasses all trajectories that are possible under the current regime, should become more complex to accommodate this. As noted above, this should result in the formation of increasingly complex structures of clusters in activity trajectory space.

If richer experience necessitates more complex activity trajectories, as well as increasingly complex structures of clusters in the space of activity trajectories, these two facets of the complexity of activity must be coupled: the subtler the discernments (differentiation in the phenomenal field) that arise from the representation of one's surroundings, or mental content in general – which is manifested as enhanced clustering in trajectory space – the richer the experience, and consequently the complexity of activity trajectories. But the converse must be true as well: as activity trajectories grow more complex, so must experience, and with the richness of experience the distinctions that are immanent in it, and hence the complexity of the realized conceptual domains. We therefore define the representational capacity of a space of trajectories as the joint (tightly coupled) complexity of (i) the structure of individual trajectories in it and (ii) the structure of the space of trajectories itself.

To move from these general considerations to operational terms, let us first consider how the complexity of the structure of a space (such as a space of trajectories), that is, configurations of clusters, can be measured. As noted above, a reasonable measure of complexity will be sensitive not only to the degree of clustering found within a space, but also to the effective dimensionality of the various configurations of clusters to be found within that space. So in essence what we would like to be able to do is simply count configurations of clusters according to their effective dimensionality.

It turns out that exactly this information, namely, the number of configurations of clusters according to dimension as a function of scale, is readily computable by the *multi-scale homology* of a space (see Fekete et al. 2009 for technical details).

In comparison to clusters, measuring the complexity of trajectories is a much more straightforward affair. Recall that our considerations led us to realize that the complexity of activity trajectories is an invariant, given an experiential state. Available evidence suggests that suitable invariants have to do with the spatiotemporal organization of activity (Makarenko et al. 1997; Contreras & Llinas 2001; Leznik et al. 2002; Cao et al. 2007; Fekete et al. 2009). In other words, activity trajectories can be classified according to experiential state: a classifying function, which we will refer to as a *state indicator function*, can be defined on activity trajectories

(i.e. over the space of activity trajectories). A state indicator function assigns each trajectory a number³ so that a given state of consciousness is associated with a typical or *characteristic value*.

This brings us to the crux of the matter: if constructed properly, a state indicator function provides a means for measuring representational capacity. As just noted, the characteristic value of a state indicator function would pick out all activity trajectories in a given experiential state, as ex hypothesi they share the same degree of complexity. In other words, it would single out the entire subspace of activity trajectories associated with an experiential state. In technical terms, this amounts to saying that the *level sets*⁴ of a state indicator function carve out experiential state-dependent spaces of activity trajectories. As these are well defined mathematical objects, their complexity, as measured by their multi-scale homology, can be computed exactly. In other words, a state indicator function provides a handle on the otherwise elusive concept of the space of all possible trajectories, and therefore on the space of possible experiences for a given system.

Note, that a complexity measure also establishes an ordering over the space of systems by their representational capacity, thereby also ruling out some classes of systems as non-conscious. To reiterate, systems that give rise to homogeneous (topologically simple) activity trajectory spaces lack consciousness altogether. That said, it is important to stress that by no means are we implying that the structure of a trajectory space alone suffices to realize experience. Rather, only activity trajectory spaces that are parceled into non-trivial level sets by a plausible complexity measure fit the bill.

We see then that the structure of the activity trajectory space is the footprint of experience, and moreover that this structure can only be understood (and quantified) from the perspective of a plausible complexity measure. If we return to the analysis of the total machine state dynamics of a digital simulation we see that it does not realize the same activity trajectory space of the brain it simulates: even if partial states are considered under numeric interpretation, the realized trajectory space could have fundamentally different topological properties resulting from implementation details. If a numeric attribution is withheld we see that so called coordinate functions have properties that drastically differ in time compared to brain ones (e.g. a two's complement system would give rise to gross discontinuities

3. Or a low dimensional vector; cf. Hobson et. al. (2000).

4. For a state indicator function $SIF: A \rightarrow \mathfrak{R}$, the level set associated with a value $c \in SIF(A)$ is the entire set of activity trajectories $a \in A$ for which $SIF(a) = c$, or all the activity trajectories that a state indicator function would assign the same score (value) – that is, they would exhibit the same degree of complexity (as measured by the SIF).

resulting from change in sign). And if the actual spatial configuration of variables is taken into account we see further discontinuities (resulting from auxiliary variable and memory management). Further still, it is hard to see how intrinsic multiscale structure can be attributed to the dynamics if even the fundamental level – i.e. numeric interpretation – is decidedly extrinsic, a fact compounded given the abovementioned spatiotemporal discrepancies as well as oddities caused by various housekeeping necessities (e.g. duplication of variables, memory management, and optimization).

The trajectory space of a digital computer lacks structure, while the trajectory space of brain dynamics has rich hierarchical structure in conscious states. Our preceding analysis shows that a digital simulation does not realize the trajectory space of the original dynamical system, and hence that such a simulation cannot be indistinguishable from the original with respect to a function such as consciousness. There is, however, still a possibility that such a simulation realizes *some* non-trivial trajectory space, and thus perhaps gives rise to some unknown form of consciousness.

To address this possibility, we need a complexity measure that would quantify the amount and the kind of structure in a space of trajectories (thereby distinguishing, for instance, trivial trajectory spaces from non-trivial ones), and would do so in a manner that is intrinsic to the dynamical system in question – that is, without resort to an external interpretation, of the kind that is part and parcel of programmable digital computer architectures and of the algorithms that such architectures support.

In the case of a digital computer, as discussed earlier, the most natural description is a discrete-state formal one, whose parts and possible states are enumerated by the machine (hardware) specifications, and whose pattern of causal interactions is dictated by the architecture (including the CPU with all its registers, the various kinds of memory, peripherals, etc.). Given this fact, an algorithm whose realization the computer runs is far from being the causal nexus of the series of events that unfolds; rather, it is merely a part of the total machine state (as it is encoded in bits just like numeric variables). For the exact same reasons, a change in the algorithm is nothing more than a kind of input: it is simply yet another external event (from the point of view of the machine) leading to a change of the current total machine state, which in turn has an effect on the consequent sequence of states.

Therefore it appears that to gauge the representational capacity of digital computers we need to consider the space of sequences of states, which would result from running every possible program, be it clever, arbitrary or even ill defined. How would this space be parceled under a reasonable complexity measure? Given stretches of machine dynamics leading to similar computational load result in

generic trajectories whose complexity is typical of the computer and not the algorithm. Such structure results from a host of subprocesses such as memory management and various optimization steps and so forth. As a whole, it is hard to see why this space – which is populated by various senseless “programs” (state switch sequences) – would not occupy some undifferentiated convex mass within the “hypercube corner” space that embeds the total machine dynamics, or at least some simply connected manifold, as is the case with various constrained physical systems (see appendix B, Fekete & Edelman 2011).

It’s about time. As the preceding arguments suggest, digital simulation must fail to give rise to consciousness because of three major shortcomings: (1) discreteness – the finite-state core architecture of digital computers leads to incapacity to represent integers, let alone real numbers, in an intrinsic manner; (2) simulation is inherently incapable of realizing the dynamical system that it is supposed to, because there is no intrinsic distinction between the part of the total causal pattern that encodes the simulated dynamical system (equations) and the part that corresponds to the system variables; (3) simulation is incapable of realizing intrinsically the appropriate multiscale organization. Together, these factors preclude digital simulation from attaining significant representational capacity – that is, generating an intrinsically hierarchical complex space of trajectories (spatiotemporal patterns) – and hence from giving rise to experience.

This conclusion does not rule out the possibility that other types of machines might be up to the task. Such machines would have to be open (i.e. engaging in input and output, at least potentially⁵), analog (possessing continuous parameters and graded states) dynamical systems that would instantiate the necessary dynamics, rather than attempting to simulate them. Our analysis so far has, however, been neutral with regard to a fundamental question, namely, whether or not such systems must be time-continuous. To engage with this question, we need to analyze the dynamics of a hypothetical machine that meets all the above requirements, yet is time-discrete. As the preceding discussion illustrates, this kind of analysis cannot ignore implementation details; accordingly, we shall conduct it by continuing with the same example that we have been using so far in examining implementation issues, a digital computer, focusing now on its state switching as a model of implementing discrete time.

5. That is, the formal scheme should be one that can accommodate a continuous stream of inputs (Hotton & Yoshimi 2011), unlike, for instance, a Turing machine. This is a necessary requirement for a formalism that purports to model brain activity, as brains most certainly are such systems.

Time in digital computers is discrete. In a typical digital computer, machine states are switched at the ticks of a central clock. This is imperative for the correct operation of most computers of contemporary design.⁶ The interval between ticks can be varied from as short as the hardware can support (a small fraction of a nanosecond in modern architectures) to years, without affecting the computational integrity of the algorithm that the machine is running. If the interval between the clock ticks is long enough for the voltage-switching transients to fade, the machine resides in a state of suspended animation for a significant, or even predominant, proportion of the total duration of a run.

Achilles and the tortoise yet again. In comparison, most stretches of experience feel continuous. For a discrete system to give rise to continuous experience, the isomorphism between the mental domain and the physical domain would have to be violated: one domain possesses essential formal properties that the other does not. One way to try and defuse this objection is to argue that the continuous aspect of experience is illusory – some sort of epiphenomenal mental paint. This line of thought leads, however, into an explanatory dead end. While experience is private and first-person, it also affords a third-person description – namely, its formal (and hence also logical) structure.

Denying this premise amounts to placing experience forever outside the realm of scientific explanation. Claiming that some aspects of experience result from structural correspondence between mental processes and physical dynamics of the medium in which it is realized, while calling for other formal properties of the underlying medium to be overlooked, is a self-defeating stance. As before, the question that arises is, what intrinsic criteria would distinguish the relevant properties from the epiphenomenal ones? We remark that isomorphism (writ large) is the only explanatory tool available to science in its engagement not only with mental phenomena, but with physical phenomena in general (e.g. Dennett 1996), making an exception for consciousness would pull the rug from under any scientific approach to it.

Experience in fits and starts. As just noted, in digital state switching transient phases are interspersed with suspended animation. Clearly, the latter phases lack representational capacity as per definition they lack spatiotemporal complexity, and hence during such phases the system would fail to give rise to experience. The complementary point is unsurprisingly, that whatever consciousness our system gives rise to is contained to the transient state switching phases. If so, we

6. In contradistinction to asynchronous digital circuits, whose domain of applicability has been traditionally much narrower.

could have no reason to believe the contents of our experience – not only could a malevolent demon fool us with regards to the outer world, but even the contents of our transient experience would be forever suspect. At any point during the static phases of the dynamics, the underlying states could be tampered with. Even if such tampering is a messy affair, as the content of each segment is independent *ex hypothesi*, we would not be able to recollect that this had happened – for that we would need to be let to undergo the structural changes necessary for forming memories, *and* instantiate those memories by undergoing through a pertinent sequence of switches.

Worse yet, a fundamental constituent of cognition and perception at large is the ability to carry out comparisons. In this scenario all comparisons would be virtual – we could experience the result of a comparison without having carried it out by arranging the underlying state accordingly. Of course, an easy objection would simply be to argue that exact manipulation to the necessary extent is simply not feasible, hence at best “metaphysically possible”. We wholeheartedly agree, and accordingly happily proceed to apply the same logic to the idea of digital state switching as sufficient for realizing phenomenology.

Digital state switching makes for a rough ride: If our system is to be a functional isomorph of a human mind, if we omit the stutter and leave only the steps in our dynamics, the joined segments would form dynamical sequences such that the realized “corrected” trajectory space is isomorphic (isometric actually) to the trajectory space realized by the human brain. However, for that to be possible it would be necessary to exert perfect control of the states of our system. Now, the dynamics of brain-like systems are described by differential equations that have time constants – parameters governing the temporal agility of the system. Thus, to enforce such a system to halt would require the control mechanism to halt the “momentum” of the system, which would lead to dampened oscillations (or at least brief transients) on the same temporal order of the dynamics (due to the exact same time constants).

If according to our story there is unique momentary experience associated with the transient ΔX (X being the total machine state), then the same must be true of $\pm a\Delta X$ ($a < 1$) (the spaces $\{a\Delta X\}$ $\{-a\Delta X\}$ obviously have the same geometry, and a geometry quite similar (depending on a) to the space of transients at large). Thus in any quasi-realistic scenario, experience would no longer be smooth, but would in fact “shimmer”, “waver” or simply oscillate.

In summary, discrete time would seem insufficient as a substrate of experience, and while the metaphysical possibility remains, we do not find that particularly disconcerting: this predicament is shared by all physical laws, which are contingent by nature.

5. Conclusion

Our analysis unearthed several inherent shortcomings of the principle of organizational invariance. First, it relies on a notion of grain that, as we have seen, is not tenable. However, without it OI is not only merely a very general statement expressing the idea that physical causes are at the root of consciousness, but by the same token results in a notion of functional equivalence of systems that applies only to identical systems. Thus, OI seems to be of little practical or theoretic use. Second, the OI principle hinges upon the wrong notion of abstraction, namely that of input/output matching. Among the dangers it harbors are the blurring between causal patterns and actual states and promoting the risk of extrinsic definitions enforcing preconceived order where in fact it does not exist. As result, OI fails to establish the thesis it was wrought to defend, namely that of machine consciousness.

All is not lost though. If the pertinent notion of abstraction is grounded in the structure of the total system trajectory space, it can be seen that while digital computers fall short as experiential machines, another class of machines – namely, open, analog, time-continuous dynamical systems – might be up for the task, provided that they are endowed with sufficient representational capacity. Regarding systems from the perspective of the structure of their possible activity trajectory space goes a long way toward remedying many of the shortcomings of OI as well as offering other theoretical benefits.

First, under this perspective simulation (whether digital or not) and realization are seen to be fundamentally different – a system and its simulation are necessarily distinct due to the need to simulate the causal nexus. Second, systems are thus seen as points within a space of systems,⁷ in which similarity, growth and degeneration can be naturally defined. Further still, system space is naturally ordered by representational capacity, enabling classification of system types (e.g. human brains, snail brains, nervous systems, and so on), thus casting

7. The space of systems from our perspective is embedded in a(n ideal) measurement space. Each point in this space would be a system's possible trajectory space. When comparing similar systems (e.g. brains), measurement (ideally) achieves the necessary standardization to compare systems. If one wishes to analyze system space in general, then it has to be standardized – i.e. made invariant to rotation, scaling, and translation. The resulting space would be a shape space, of the kind studied in shape theory (e.g. Le & Kendall 1993). Note, however, that complexity measures such as multi-scale homology are invariant under simple transformations and thus achieve an ordering (and metric) even on non-standardized spaces.

the notion of functional equivalence (or rather functional similarity) in more concrete terms.

When it comes to the more ambitious part of the OI thesis, namely similarity of experience of machines of different classes, things become more complicated. The structure of the possible trajectory space is the physical counterpart of the perceptual/conceptual system that a system's dynamics gives rise to. In the case of organisms capable of rich experience, such as humans, this necessitates that this space possess hierarchical (multiscale) intrinsic cluster structure expressing the distinctions found in experience. Thus, for example, if empirical studies show that such structures require n levels of organization, it would make sense to suggest that even if systems are not equivalent at the lowest level, they can still share equivalent minds as long as they are equivalent at the remaining higher levels.

This claim of equivalence would be tenable at least as far as the third-person attributes of experience are concerned, as ipso facto both beings in question would not only share the same conceptual/perceptual system, but the same thought processes. At first blush, it would seem that this theoretical move is open to slippery-slope counter-arguments. Yet from a metric perspective, classes of systems (and by the same token levels of organization) form clusters – that is, categories – in system space. As is always the case with categories in a metric space, they will be fuzzy at the borders, and illustrative to the extent that exemplary members of classes (i.e. those that are situated well within the cluster) are considered. In any case, this stance on the issue of equivalence does appeal in some sense to the idea of individuating experience through behavior. As such, it would certainly fail to impress some theorists – a predicament that we are all probably stuck with, given that another person's experience is to us fundamentally inaccessible.

Acknowledgements

TF wishes to thank Yoav Fekete for extremely insightful discussions of various computer science core issues.

References

- Alvarez-Maubecin, V., García-Hernández, F., Williams, J.T. & Van Bockstaele, E.J. (2000). Functional coupling between neurons and glia. *The Journal of Neuroscience*, 20(11), 4091.
- Bennett, M.V.L. & Zukin, R.S. (2004). Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron*, 41(4): 495–511.
- Block, N. (1980). Troubles with functionalism. *Readings in Philosophy of Psychology*, 1, 268–305.

- Cao, Y., Cai, Z., Shen, E., Shen, W., Chen, X., Gu, F. & Shou, T. (2007). Quantitative analysis of brain optical images with 2D C0 complexity measure. *Journal of Neuroscience Methods* 159(1), 181–186.
- Chalmers, D.J. (1995). Absent qualia, fading qualia, dancing qualia. *Conscious Experience*, 309–328.
- Connors, B.W. & Long, M.A. (2004). Electrical synapses in the mammalian brain. *Annu Rev Neurosci*, 27, 393–418.
- Contreras, D. & Llinas, R. (2001). Voltage-sensitive dye imaging of neocortical spatiotemporal dynamics to afferent activation frequency. *Journal of Neuroscience*, 21(23), 9403–9413.
- Dennett, D.C. (1996). *Darwin's dangerous idea: Evolution and the meanings of life*, Simon and Schuster.
- Eroglu, C. & Barres, B.A. (2010). Regulation of synaptic connectivity by glia. *Nature*, 468(7321), 223–231.
- Fekete, T. (2010). Representational systems. *Minds and Machines*, 20(1), 69–101.
- Fekete, T. & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*.
- Fekete, T., Pitowsky, I., Grinvald, A. & Omer, D.B. (2009). Arousal increases the representational capacity of cortical tissue. *J Comput Neurosci*, 27(2), 211–227.
- Fields, R.D. & Stevens, B. (2000). ATP: An extracellular signaling molecule between neurons and glia. *Trends in Neurosciences*, 23(12), 625–633.
- Hebb, D.O. (1988). *The organization of behavior*, MIT Press.
- Hotton, S. & Yoshimi, J. (2011). Extending dynamical systems theory to model embodied cognition. *Cognitive Science*.
- Le, H. & Kendall, D.G. (1993). The Riemannian structure of Euclidean shape spaces: a novel environment for statistics. *The Annals of Statistics*, 1225–1271.
- Leznik, E., Makarenko, V. & Llinas, R. (2002). Electrotonically mediated oscillatory patterns in neuronal ensembles: An in vitro voltage-dependent dye-imaging study in the inferior olive. *Journal of Neuroscience*, 22(7), 2804–2815.
- Makarenko, V., Welsh, J., Lang, E. & Llinás, R. (1997). A new approach to the analysis of multidimensional neuronal activity: Markov random fields. *Neural Networks*, 10(5), 785–789.
- Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30(1), 63–80.
- Milner, P.M. (1974). A model for visual shape recognition. *Psychological Review*, 81(6), 521.
- Pannasch, U., Vargová, L., Reingruber, J., Ezan, P., Holcman, D., Giaume, C., Syková, E. & Rouach, N. (2011). Astroglial networks scale synaptic activity and plasticity. *Proceedings of the National Academy of Sciences*, 108(20), 8467.
- Pylyshyn, Z.W. (1980). The 'causal power' of machines. *Behavioral and Brain Sciences*, 3(03), 442–444.
- Ricci, G., Volpi, L., Pasquali, L., Petrozzi, L. & Siciliano, G. (2009). Astrocyte–neuron interactions in neurological disorders. *Journal of Biological Physics*, 35(4), 317–336.
- Rosenthal, D.M. (2005). *Consciousness and mind*, Oxford University Press, USA.
- Scemes, E. & Giaume, C. (2006). Astrocyte calcium waves: What they are and what they do. *Glia* 54(7), 716–725.
- Shagrir, O. (2006). Why we view the brain as a computer. *Synthese* 153(3), 393–416.
- Shigetomi, E., Bowser, D.N., Sofroniew, M.V. & Khakh, B.S. (2008). Two forms of astrocyte calcium excitability have distinct effects on NMDA receptor-mediated slow inward currents in pyramidal neurons. *J Neurosci*, 28(26), 6659–6663.

- Theodosis, D.T., Poulain, D.A. & Oliet, S.H.R. (2008). Activity-dependent structural and functional plasticity of astrocyte-neuron interactions. *Physiological Reviews*, 88(3): 983.
- Von Der Malsburg, C. (1994). The correlation theory of brain function. *Models of neural networks II: Temporal aspects of coding and information processing in biological systems*, 95–119.
- Wang, Y., Barakat, A. & Zhou, H. (2010). Electrotonic coupling between pyramidal neurons in the neocortex. *PLoS ONE* 5(4), e10253.
- Wittgenstein, L. (1953). *Philosophical investigations*. New York, Macmillan.

Restless minds, wandering brains

Cees van Leeuwen & Dirk J.A. Smit

RIKEN BSI, Japan and KU Leuven, Belgium / VU University Amsterdam,
the Netherlands

1. Introduction

In “The restless mind”, Smallwood & Schooler (2006) describe *mind wandering* as follows: “the executive components of attention appear to shift away from the primary task, leading to failures in task performance and superficial representations of the external environment” (p. 946). Characteristically, mind wandering is seen as distractedness; a shift of attention toward internal information, such as memories, takes resources away from the task; this leads to less accurate awareness of external information and potentially a failure to achieve the goal of task – thus, mind-wandering is tantamount to disfunctionality.

Here we will make a case for a more positive view of mind wandering as a possible important element of brain function. But first, let us distance ourselves from introspective reports; as our mind wanders, we are often unaware of the contents of our current experiences (Schooler 2002). This means not only that mind-wandering is underreported, but also that it is likely to remain undetected until something goes wrong. The claim that mind-wandering is dysfunctional, therefore, may largely be a matter of sampling bias.

We propose to use psychophysical methods instead to study mind-wandering. Whereas introspective reports are often unreliable, extremely reliable reports on experience can be obtained in psychophysics. This will allow us to investigate what the antecedent conditions are for mind-wandering as a cognitive phenomenon, what possible positive effects it may have, and how individuals may differ in their mind-wandering brains.

The psychophysical approach may be applied to cases somewhat like the following. Study Figure 1 for a while and you will repeatedly experience spontaneous changes in the grouping of its components, to which we sometimes, but not always, attribute meanings: a star, a staircase, an open cardboard box, a toy house, etc. This phenomenon is known as perceptual multi-stability. Some

instances of perceptual multi-stability may be more compelling than this one. The present demonstration, however, has an important feature that we may also expect from the real world: it is suggestive of a rich variety of different structures, without supporting any of these unambiguously. In the real world, this is due, for instance, to figures occluding each other or, over time, to partial views of the same image (cf. Peterson & Hochberg 1983).

The perceptual mind-wandering exemplified by Figure 1 might perhaps seem dysfunctional – isn't our visual system supposed to insure that our perceptions are robust and stable? True: evolution has assigned to the visual system the role of providing *tolerance against variability in the visual world*: to detect recurrent patterns in continuously shifting and changing stimulation. Recurrent patterns are detected at a range of levels in the perceptual process: early, mid, and high-level vision. Early vision provides tolerance against variation in, for instance, contrast and color under variable illumination (e.g. Healy & Slater 1994; Lucassen & Walraven 1993). Mid-level vision involves *perceptual organization*: the grouping of early features into tentative object structures (Gepshtein & Kubovy 2007; Kubovy, Holcombe & Wagemans

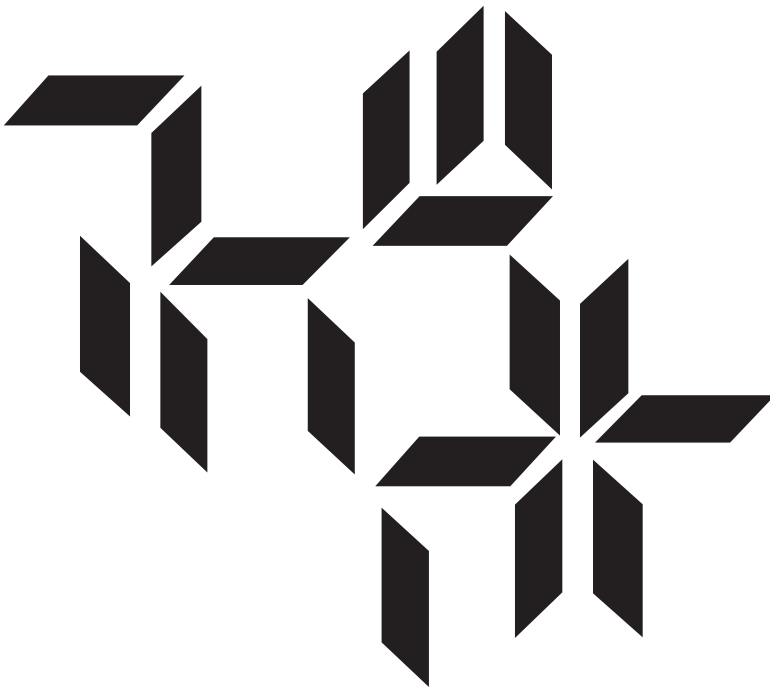


Figure 1. From van Leeuwen (2007). Perceptual multi-stability as mind-wandering. A wandering mind can discover several structures in this figure, such as the star (lower right part of the figure), a staircase on the main diagonal, a house (in the upper right), etc.

1998). The groupings are unaffected by a variety of transformations, involving repetition, rotation and reflection (Garner 1962), and scaling. High-level vision renders them as identifiable objects despite, for instance, occlusion and location or viewpoint changes (Thier, Haarmeier, Chakraborty et al. 2001).

Even though perceptual mind wandering threatens the stability of representations, it may for the exact same reason have an adaptive function. While detecting recurrent patterns is a major factor in perception, the visual system must also satisfy the evolutionary demands of *flexibility*: it must enable rapid revision if the situation requires so, in order to respond to unexpected objects, events, or to discover new aspects to familiar ones.

Multistability is a sign that flexibility is at work in our perceptual system. When the perceptual system groups together the components of the “star” pattern in Figure 1, the “staircase” pattern can no longer be formed, as some of its components are already assigned to the flower. Mind-wandering could break up the existing pattern structure, such that the components can be reassigned to a different pattern, like the staircase. Items in the background – or, if you like, outside of the focus of our attention – tend to receive weaker groupings than those in the foreground (Rock & Gutman 1981). Thus, when the mind wanders away from a certain grouping, it will dissolve and release its components for novel use. This means that perceptual mind wandering allows the perceiver to detect initially hidden *surplus structure*.

The detection of surplus structure plays an important role in evolution. It allows a perceiver to discover something in the current scenery that is not detectable at first sight, such as a camouflaged predator hiding in the bushes. It also characterizes some highly cultured aspects of perception. Discovering the hidden structure in a visual display plays a role in the creative process of designers (Verstijnen, van Leeuwen, Goldschmidt et al. 1998a,b; Verstijnen, van Leeuwen & Hamel 2000) and artists (van Leeuwen, Verstijnen & Hekkert 1999), who typically use externalizations of their creative process, such as sketches, to detect the surplus structure in their images and guide the further course of their design process. Towards the end of the design process, the various structures are then assembled to be integrated in the final design (Jaarsveld & van Leeuwen 2005).

This last aspect is particularly intriguing, as the notion of creativity has been associated with the production of novel structure, but not any novel products are equally creative. Many of the artists we investigated stated that they were trying to achieve some form of “balance” in their design – a property they were unable to describe further. Designers and artist, it seems, tacitly know how to lure perceivers into looking longer their work, by offering them the pleasure of discovering over time the surplus structure of the work on display.

These mechanisms are interesting for the following reason: It is still a major puzzle how perceptual systems accommodate the seemingly opposing demands

imposed by evolution of robustness under variability versus flexibility. Most theoretical models, including recurrent neural networks, are geared towards stability. Their behavior is robust, but not flexible. Hopfield (1982), for instance, defined a function describing the “energy” of a representational state, and proposed that approaching an attractor from initial conditions given by stimulation corresponds to minimizing the system’s energy expenditure. Models like this are inherently oriented towards stability. They do not change their representation, unless perturbed by, for instance, newly incoming information. Such systems would not be capable of discovering surplus structure.

1.1 Escaping robustness

We may consider, as Hopfield did, to use classical Gaussian noise for escaping an attractor. The concept of an internal Gaussian noise source attenuating the signal in sensory channels has gained wide acceptance in the study of perceptual processes due to signal detection theory (Green & Swets 1966). However, for driving change in a stability-oriented system it is highly inefficient; the very stability of attractors protects them against change. We might, alternatively, consider systems far from equilibrium (Maturana & Varela 1980), in which a small perturbation may have large consequences (the “butterfly effect”). But this would clash with the requirement that perceptual representations should be stable. Far-from-equilibrium systems are flexible but not robust.

Novel approaches are emerging from multidisciplinary collaborations in the physical, biological, computational, and social sciences, operating under the banner of complex systems science. Complex systems are, in principle, any network of multiple components that interact nonlinearly, giving rise to emergent behavior. Complex adaptive systems in addition have the capacity to adapt to their environments and so achieve optimality in their functioning. Examples include social insect and ant colonies, the biosphere and the ecosystem, the immune system, large-scale online systems and, of course, the brain.

In the current text, we purport to show, first, that the dynamics of perceptual mind wandering bears the hallmark of a complex adaptive system. Next, we will discuss what function it may have: mind-wandering, we propose, is for integrating perceptual information with its context. In regulating the integration, we will attribute a crucial role to alpha band activity observed in human scalp electro-encephalography (EEG; Linkenkaer-Hansen et al. 2004; palva & palva, 2007). We will discuss alpha EEG in a state prone to mind wandering, which is called the “resting state”, and observe that complexity is equally prominent there as in the perceptual dynamics. We will discuss the characteristics of the dynamics underlying the complexity; a dynamics in which stability and instability of neural

synchronization patterns are tightly interwoven. We will discuss what happens to this dynamics when mind-wandering is effectively put aside, and the observer concentrates on a task. At crucial points, we will discuss what the implications of these results are for our visual conscious experience.

2. Perceptual switching

To simplify from Figure 1, a special case of perceptual multi-stability can be found, suitable for psychophysical measurement: the class of widely-known bi-stable or ambiguous figures, such as the Necker cube (Necker 1832; see Figure 2). Prolonged exposure typically leads to switching between these interpretations. This behavior is involuntary to a large extent and occurs independently of external cues such as local disambiguation (Peterson & Gibson 1991) or internal ones such as eye-movements (Einhäuser, Martin & König 2004; Nakatani & van Leeuwen 2005, 2006; van Dam & van Ee 2006). We conclude that there are mechanisms within the visual system that go against the stability of its interpretations.

Along the lines set out by Hopfield, spontaneous switches in visual perception might simply be considered a matter of Gaussian noise. Empirical evidence suggesting that such a mechanism is responsible for perceptual switching is found in the *dwelt times* (Borsellino, De Marco, Allazetta, Rinesi & Bartolini 1972; Levelt 1967). These are the durations with which a certain interpretation is maintained. They follow a Gamma distribution, consistent with the view that they are produced by a number of independent occurrences of a chance event. This observation has motivated discrete state models of which randomly alternate between states after a critical number of random events that favor the new state (Taylor & Aldridge 1974).

A Hopfield-type of model giving rise to a Gamma-like switching behavior would have two roughly equivalent minima (a “double well” model). Noise events can drive the system a step away from a minimum, and an accumulation of such events can drive it sufficiently far away to enable approach of the other minimum. Note that the spontaneous changes are still driven by Gaussian noise and thus dwell times would still be sequentially independent. In reality, however, these correlations are consistently above zero and decrease with lag (Bassingthwaite, Liebovitch & West 1994; van Ee 2009); significant temporal correlations may survive over seconds or even tens of seconds. These sequential dependencies suggest that macroscopic fluctuations play a role in switching as well. Indeed: Kim, Grabowecky, and Suzuki (2006) induced stochastic resonance in switching by periodically alternating the stimulus, thereby demonstrating the presence of macroscopic noise in the system.

Consider, accordingly, a version of the double well model that cycles spontaneously between approach and avoidance of the equilibria; in other words between being governed by stability and flexibility. Suppose, for instance, that the well in which the system is residing becomes gradually shallower due to mechanisms such as adaptation or competition. This means that fewer noise events would suffice to drive the system out of its state. This assumption has been embedded into macroscopic models of the dynamics of switching behavior as a phase transition (Ditzinger & Haken 1989, 1990). In such models, the fast noise and a slow dynamic cycle work together to produce perceptual switching behavior.

In olfactory perception, Skarda and Freeman (1987) have described a system in which a similarly slow dynamic cycle is coordinated with the breathing rhythm; upon inhalation it is geared towards equilibrium states, and thereby ready to classify an incoming odor; upon exhalation the attractors are removed by changing the well depth. As a result, the system becomes optimally sensitive to new information. Freeman and van Dijk (1987) envisaged the same kind of mechanism for visual perception: the “perceptual sniff”.

Köhler and Wallach (1944) proposed the slow dynamics to be neural fatigue or satiation. There is no evidence of neural fatigue at the level of active configurations. There is, of course, the well-established phenomenon of neural adaptation but this takes place at the local level of ion currents conductivity in the membrane of the neuron (Sanchez-Vives, Nowak & McCormick 2000), and is therefore not sufficiently selective at the level of global perceptual patterns (Barlow & Földiák 1989). Adaptation to patterns, according to these authors, occurs through a mechanism of anti-Hebbian decoupling between cells that are simultaneously active. Such a macroscopic mechanism could be useful to explain the steady, continuous increase in switching rate with prolonged presentation of a stimulus. Correlations between subsequent dwell times could be explained by fluctuations in adaptation rate (van Ee 2009).

But there is another possibility: rather than as noise added to a system at two different levels, we may regard the system's fluctuations as fractals. Consider a series of dwell times: they fluctuate irregularly from one time to the next. When these data have fractal characteristics, the means of different subsections fluctuate in a similarly irregular manner. Such similarities across all scales are estimated by the Hurst- or H-coefficient, of self-similarity (Beran 1992; Mandelbrot & Wallis 1969) or, for non-stationary series, the DFA exponent (Peng, Havlin, Stanley et al. 1995). With a H exponents of .5 the series consists of Gaussian noise. This means that adjacent time intervals are uncorrelated (or have correlations that quickly goes to zero with lag). If the exponent is < 0.5 , the system tends to switch between high and low values and $> 0.5 < H < 1.0$ means that the signals preserve trends over a long period – a long-term memory (Wagenmakers, Grünwald & Steyvers 2006).

Along with self-similarity as in fractals, such memory is the hallmark of complexity in time series. The phenomenon is sometimes referred to as $1/f$ noise; note, however, that a $1/f$ -type spectral density is necessary but not sufficient for fractal noise. It is possible, for instance, to scramble a fractal series, while preserving its spectral characteristics. Such series are used as surrogate data in bootstrapping estimation of statistics such as H .

Recently, Gao et al. (2006) presented H estimates for dwell times in the Necker cube, clearly demonstrating their fractal nature. Gao's observations complement what is found in a large variety of behavioral tasks, such as mental rotation, lexical decision, speeded visual search, estimation of distance, estimation of rotation, estimation of force, estimation of time, simple reaction times, choice reaction times, and word naming (Gilden 1997, 2001; Gilden, Thornton & Mallon 1995; Kelly, Heathcote, Heath et al. 2001; Van Orden, Holden & Turvey 2003). It seems, therefore, that a complex systems characterization is appropriate for a much wider range of behaviors other than perceptual switching. This underlines the general relevance of perceptual instability for understanding perception, cognition, and the brain mechanisms underlying these processes.

3. An EEG-study of perceptual mind wandering

In a master thesis (Smit, Heslenfeld, Kenemans et al. unpublished manuscript), we investigated perceptual switching in a Necker cube paradigm adopted from Peterson & Hochberg (1983). These latter authors used Necker cubes that were locally disambiguated (Figure 2, left). When the disambiguated region of the picture is overtly or covertly (Peterson & Gibson 1991) attended to, observers perceive the orientation according to the local bias. Regional bias, however, fails to influence the perceived orientation when a different, unbiased region is attended to. This result shows that when the figure is presented in isolation and attention is focused on a part of it, the remaining parts are only loosely integrated with the rest of the figure, resulting in tolerance for global inconsistencies. According to the authors, this justified the much stronger claim that generally objects are perceived in a piecemeal fashion.

Smit et al. (1994) surrounded the Necker cube at a distance of 4.5° in all directions with eight wire-frame images, depicting cubes, all in one of two possible orientations. The central Necker cube was shown with two alternative local biases (indicated by the circles marked in the left hand side of Figure 2), in combination with one out of two alternative contexts as in the right hand side of Figure 2. Each unique stimulus was repeated 7 times. Participants reported on the perceived orientation of the central cube. The question was, whether adding surrounding

context to the Necker cube has any influence on its piecemeal perception and, in particular, whether perceptual mind-wandering (to the surroundings) strengthens figural integration.

During presentation of the surrounded Necker cube task-irrelevant stimuli (probes) occurred. The probes consisted of. 4° diameter black dots. They were presented for 50 ms at a variable rate of between 800 to 850 ms, left or right visual hemifield and near 1.4° or far 5.0° from the fixation point. Each Necker cube presentation allowed for the presentation of 80 probes, 20 on each location, in random order. Observers were told to ignore the probes as well as the context cubes.

From nine participants, two scores were calculated for each condition from the proportion of time the cube was reported as seen in either orientation. Preference scores (I_{pref}) were obtained as in Formula 1. They indicate in which of the two possible orientations the cube is most frequently seen. They range from -1 to $+1$. Positive I_{pref} mean a preference for the view-from-below orientation of the cubes, negative ones for the view-from-above orientation; zero means equal preference. Next, responses were analyzed in terms of I_{CoCo} , the degree of conformity of their preferred orientation with the context, as in Formula 2. A value of $+1$ stands for total consistency with the context, -1 for total inconsistency.

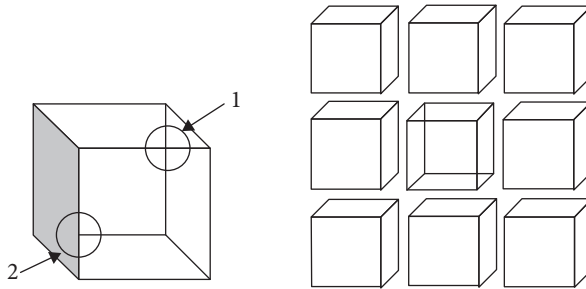


Figure 2. Left: local bias in the Necker cube. Circles indicated by arrows represent the two alternative areas to which a prior instruction draws attention. No arrows or circles were present in the actual display. Area 1 contains an X-junction indicating an ambiguous cube; Area 2 contains a T-junction which yields a strong bias to the orientation of the cube; right: global bias: a Necker cube in surrounding context

$$I_{pref} = \frac{\text{time viewed from below} - \text{time viewed from above}}{\text{time viewed from below} + \text{time viewed from above}} \quad (1)$$

$$I_{CoCo} = \frac{\text{time Consistent} - \text{time Inconsistent}}{\text{time Consistent} + \text{time Inconsistent}} \quad (2)$$

Analysis of variance on I_{pref} showed main effects of local bias ($F(1,8) = 14.57$; $MS_{err} = .203$; $p = .005$) and global bias ($F(1,8) = 8.88$; $MS_{err} = .011$; $p = .018$). No

interaction was found ($F < 1$). The preferred interpretation, therefore, is independent of the focus of attention, a result which contrasts with Peterson & Hochberg (1983) and Peterson & Gibson (1991), who found that in isolated cubes, local biases operate locally, i.e. within the focus of attention. Here, the local biases operated globally and equally across the whole figure. The mere presence of a surrounding context, even though instruction discouraged observers to attend to it, is sufficient to eliminate the piecemeal object perception effect. Observers on average prefer to be in accordance with the global context. In other words, they not only prefer consistency within the figure, but also between the figure and its surroundings. The presence of the surrounding figures has not distracted the observers, but rather it has led them to enhanced integration of the central figure and reduced tolerance for inconsistencies within it.

Table 3. Data from Smit et al. (1994). Context consistency per subject. I_{CoCo} is the index for context consistency. Observers with negative I_{CoCo} (1) and/or too uneven numbers of probes within conditions (2) were excluded from the probe ERP analyses

Observer	I_{CoCo}	Inclusion
1	.0447	+
2	-.0115	– (1,2)
3	-.0303	– (1)
4	.0663	– (2)
5	.0484	+
6	.0438	+
7	.1123	+
8	.1404	– (2)
9	.0416	+

Table 3 shows that even though the majority of individual observers tended towards percepts that were in accordance with the context, there was considerable variability in individual preference for context consistency. To study whether the context consistency was related to perceptual mind-wandering, we measured event-related potentials evoked by probe stimuli (Probe ERP) in the observer's EEG. For trials where center and surrounding were oppositely biased, episodes in which the perceived orientation of the Necker cube in the center went with the surrounding context were classified as wandering; those in which it went with the local bias in the Necker cube itself were classified as focused. The remaining trials were used as baseline. ERPs were averaged per event (probe left/right and probe far/near), per cube configuration shown during that time (context from above/ from below and stimulus from above/from below), and per perceived orientation

at the time of presentation of the probe (consistent with the surrounding context or with the local bias).

Based on the extant ERP literature and visual inspection of the grand-averaged ERP curves we selected four time windows that matched positive and negative peaks. Windows 1 and 2 represented P1 and N1, 100 ms to 150 ms and 160 ms to 230 ms post-stimulus. Windows 3 and 4 were selected to cover the late positivity. Window 3 represents 200 ms to 270 ms, and Window 4 270 ms to 350 ms. For each of these the mean amplitude was calculated. The positivity that appeared from ca. 180 ms across the midline leads showed in Time Window 4 over the frontal region the effect of Figure 3.

Figure 3 shows the average amplitude induced by near and far probes. Whereas the amplitudes showed a steep decline with “focused” conditions, those in the “wandering” condition were essentially the same across probe distance. Attention – measured with the probes – seemed evenly distributed in the mind-wandering condition, and indeed focused when observers were influenced by the local cue. The frontal activity may arise from the frontal eye fields, which are involved in directing attention (Crowne 1983). Even though electrode location does not reliably reflect the underlying EEG activity and the number of observers in this study was rather small, the results were at least suggestive of the view that context congruency depends on a wide attentional focus, and temporary increase of a mind-wandering state.

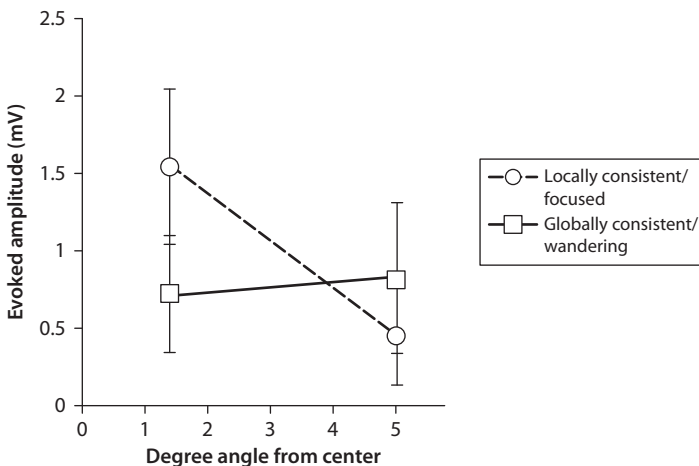


Figure 3. Data from Smit and van Leeuwen (1994; unpublished data). Probe ERPs in μV means and standard errors. The probe distance $\times$ interpretation interaction was found ($F(1,4) = 9.3$; $p < .05$) in the time window of 270–350 ms, recording site Fz. Similar trends were observed in neighboring electrodes

The degree to which individual observers differ in context consistency varied greatly. This led us to remove subjects for which the interpretation times were so skewed that reliable ERPs were impossible, thus limiting the generalizability of the ERP results presented above. However, these individual differences also provided a window to the brain mechanisms underlying mind-wandering. To understand what causes an individual's ability to focus, we additionally obtained from lead O2 a simple 3 minute eyes-closed resting EEG recording, from which we calculated the dominant alpha frequency. This frequency generally was easily determined, but when two clear peaks in power spectrum were visible, the mean of both peak frequencies was calculated. This measure ranged from 9.38 Hz to 11.72 Hz across subjects. The correlation between individual peak frequency and I_{CoCo} was found to be negative, very strong and significant ($r = -.717$; $N = 9$; $p < .05$). Slower oscillations correspond to increased context dependency.

To interpret this correlation, observe that "alpha activity is a far from unitary phenomenon, and discussions of alpha must take into account the anatomical and behavioral context." (Foxy & Snyder 2011, p. 10). Alpha frequency was found to increase during execution of difficult tasks compared with more simple ones (complex addition and mental rotation vs. simple addition and visual imagery). The increase is largest in the hemisphere that is dominant for a specific task, i.e. arithmetical tasks for the left, and visuo-spatial tasks for the right hemisphere (Osaka 1984). Alpha frequency also correlates positively with specific verbal and non-verbal abilities (Anokhin & Vogel 1996; Jaušovec & Jaušovec 2000; Shaw 2004), but a relationship with general intelligence is contentious; some have reported a correlation (Angelakis, Lubar, Stathopoulou et al. 2004; Doppelmayr, Klimesch, Stadler et al. 2002) others find inconsistent (Gaser, Von Lucadou-Muller, I., Verleger et al. 1983) or no effects (Posthuma, Neale, Boomsma et al. 2001; See Smit, Wright, Hansell et al. (2006) for a review).

Cortico-cortical interactions involved in alpha activity have a function in biasing attentional processes. Whereas enhancement of neural processing at the focus of attention is modulated by gamma frequency (Gruber, Müller, Keil et al. 1999; Sokolov, Lutzenberger, Pavlova et al., 1991), alpha activity is involved in suppressing the attention of neurons responsible for processing stimuli outside of the focus of attention (Lopes da Silva 1991). Alpha oscillations may thus represent a certain rhythm of "pulsed inhibition" (Mathewson, Lleras, Beck et al. 2011) on attentional processes. From this we hypothesize that alpha frequency reflects the pulsation: the higher the alpha frequency, the faster the process is updated. What are the implications of this hypothetical mechanism with regard to mind-wandering?

Recall that mind wandering does not inevitably lead to reduced cognitive functioning, as the association of mind wandering with a state of distractibility would suggest (Smallwood & Schooler 2006). Mind-wandering yields a positive contribution to perception. The more mind-wandering, the more likely that interpretation is taking context into account. Perceptual mind-wandering is beneficial to the perception of a figure, as it helps maintaining a consistent image of the whole – the figure is no longer piecemeal, i.e. broken into elements by an overly focused attention (Hochberg & Gibson 1991); the figure itself is seen in relation to its surrounding. In a way, this is a natural mode of perception. One of the functions of perception is to give a holistic, integral account of what we have in front of us. It is no less natural, however, than the analytic, context-independent view; both have their own preferred domains (Lachmann & van Leeuwen 2008; van Leeuwen & Lachmann 2004).

Individuals differ with respect to their preferred perceptual style. This difference has been addressed with the notion of field-(in)dependence (Witkin 1950). Here we observe holistic/analytic processing and field-(in)dependence to depend on a common mechanism: the pulsed inhibition of attention, the rate of which is controlled by the peak frequency in the alpha rhythm.

The peak frequency operates as a pulsed inhibition of attention; a “reset” on the contextual information. The less frequent these resets, the more contextual information percolates into the mental representation. Down in the early visual areas, which preserve the retinotopic organization of the visual image, this would translate to a narrower spatial focus of attention. Analytic and holistic processing both involve attentional selection (Boenke et al. 2009). With analytic processing comes a style of spatially focused attention. With holistic processing, attention is “spilling over” to neighboring regions (Lavie & Cox 1997). The latter is also the preferred style of the field-dependent.

4. The dynamics of mind-wandering

So far, we have proposed that (perceptual) mind wandering is, at least sometimes, an active, cognitively relevant process, and that the instabilities it produces signify complexity. Can we observe these effects in brain activity? The results mentioned above found effects in the alpha range of EEG activity. It has been claimed that alpha (9–11 Hz) and beta (15–30 Hz) decrease during mind wandering whereas theta (4–7 Hz) and delta (2–3.5 Hz) EEG activity increases (Braboszcz & Delorme 2011). But this merely indicates that mind wandering was observed under conditions of low vigilance, and may therefore be subject to the sampling bias noted in the

beginning of this text. Since we wish to refrain from introspective reports, the best approach to study the dynamics of mind wandering is to look at the spontaneous EEG in the resting state – an awake neutral state that is not associated with any specific task and that is prone to mind wandering. We were interested in whether the EEG, and hence the wandering, could be understood as a classical random walk or rather, as perceptual switching research suggests, as a process of a complex system.

Pioneering work by Lehmann and coworkers showed that certain patterns in the spatial distribution of the potential across the scalp are systematically preferred (Lehmann, Ozaki & Pal 1987). The authors discovered typical concentric potential distributions around two extrema (maximum and minimum), which remained stationary for some time, and then suddenly jumped to new locations. They considered each such period as reflecting the type and content of information, which was momentarily processed in the brain. They called these periods ‘brain microstates’. More recently, the $1/f$ scaling behavior characteristic of complexity has been observed in amplitude fluctuations of 10 and 20 Hz oscillation (Linkenkaer-Hansen et al. 2001).

Important processes, such as phase synchronization between EEG signals or traveling waves in EEG potentials, however, are not necessarily accompanied by potential extrema. We therefore turned to dynamic phase synchrony as our analysis tool of choice. Phase synchronization of EEG has widely been used as a tool to study dynamic interactions among brain areas (Fell, Klaver, Lehnertz et al. 2001; Sarnthein et al. 1998; Silberstein, Danieli & Nunez 2003; Bhattacharya, Petsche, Feldmann et al. 2001; Schack, Weiss & Rappelsberger 2003) and is considered to reflect long-range neural integration (Varela, Lachaux, Rodriguez et al. 2001) and interaction in the brain (Breakspear, Williams & Stam 2004; Freeman & Rogers 2002; Nakatani, Ito, Nikolaev et al. 2005).

Ito, Nikolaev & van Leeuwen (2005) observed flexible, dynamic synchronization patterns in human scalp-recorded brain activity (EEG) in resting state EEG. The spatial structure of these patterns reveals a high coherence between frontal and occipital regions, and lower coherence between central and other regions (Thatcher, Krause & Hrybyk 1986; Ozaki and Suzuki 1987). As for the temporal structure: when phase synchronization over the entire scalp is strong, phase patterns took either of two forms; one is a gradual phase shift, observed between frontal and occipital regions; a traveling wave of electrocortical activity of which the direction of propagation is predominantly from anterior to posterior in three out of four participants, and opposite in the remaining one (Figure 4). The other pattern is a stepwise pattern with a sudden phase shift in the central region. This pattern may correspond to a standing wave composed of two traveling waves propagating in opposite directions.

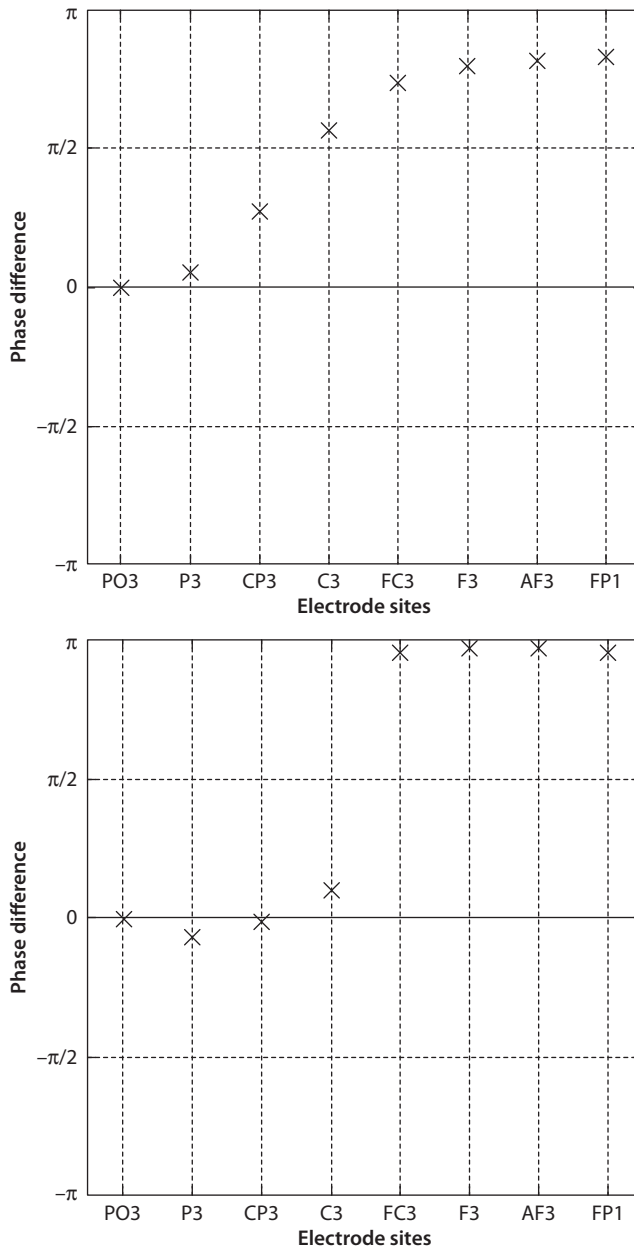


Figure 4. From Ito et al. (2005). Snapshots of two typical phase patterns during phase synchronization. The phase difference is shown as the relative phase at each site in relation to the phase at the PO3 site. Top: Phase pattern with a gradual shift (pattern A). Bottom: Phase pattern with a sudden shift (pattern B)

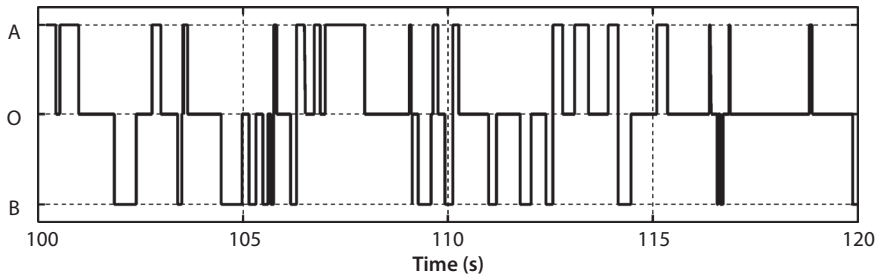


Figure 5. From Ito et al. (2005). A typical segment of time interval series of phase patterns. Taken from data of participant FEB22. The graph is above or below O when the phase pattern is classified as pattern A or B. Most often during the period when the graph is at O, the phase pattern is unstable and changing over time

Previous studies have resulted in conflicting reports on whether traveling waves predominantly occur from frontal to occipital (Shaw & McLachlan 1968; Inouye, Shinosaki, Toi et al. 1995) or from occipital to frontal (Nunez 1974; Petsche & Sterc 1968). Ito observed both to occur at different times, and sometimes simultaneously and superimposed. Most importantly, the patterns show an alternation highly akin to perceptual switching (Figure 5). The time periods during which a wave pattern remains apparently stable are referred to as *laminar periods*.

The laminar periods show $1/f$ scaling over two orders of magnitude in their durations. Gong, Nikolaev & van Leeuwen (2003) observed that the growth rates (I_k/I_{k+1}) of laminar periods I_k have $1/f$ characteristics over 3 orders of magnitude. Fluctuations in intervals of spontaneous synchronization, therefore, have no characteristic preference for interval durations, once the common average is removed (Figure 6).

Van de Ville et al. (2010) found long-range dependency, which is eliminated from the data when the durations of the intervals of synchrony are artificially equalized. This indicates that interval duration is a key characteristic of the dynamics. The observed $1/f$ signature and the long-term dependency are in accordance with the fractal characteristics of the time series. Patterns of the fast time scale recur in the slow time scale. This implies that patterns observed in rest-state EEG correlate with those of the BOLD signal in fMRI (Laufs et al. 2006; Mantini et al. 2007).

In an effort to characterize the underlying dynamics of spontaneous activity, Ito, Nikolaev & van Leeuwen (2007) studied the whole-head two-dimensional pattern of pair-wise phase synchronization patterns, using high density EEG. In Figure 7, a transition between two relatively stable traveling wave patterns can be observed. The stability of wave patterns was investigated and its fluctuations tested against randomness using surrogate series (Figure 8).

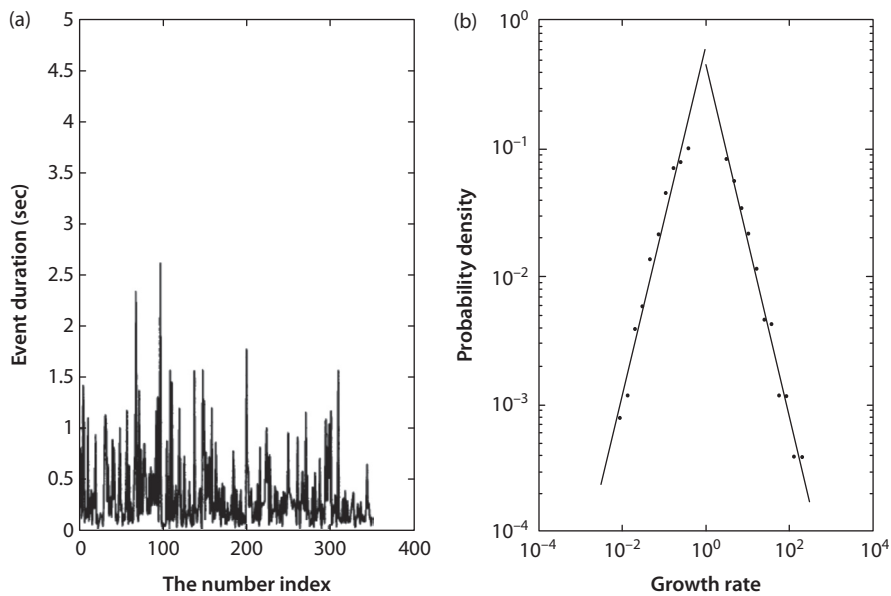


Figure 6. From Gong et al. (2003). (a) Duration of dynamical phase synchronization episodes versus their index number of the events; (b) scaling of growth fluctuation in the duration of dynamical synchronization

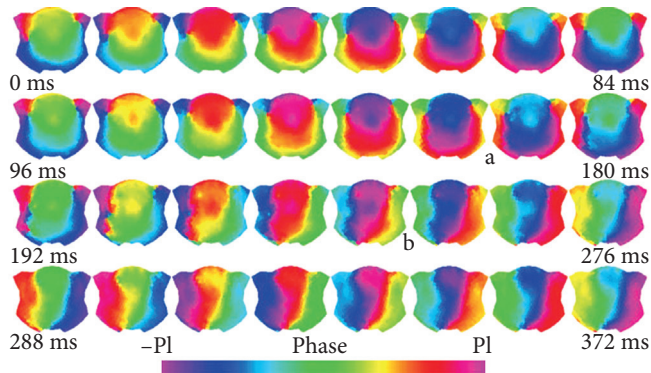


Figure 7. From Ito et al. (2007). Typical phase map time series, containing snapshots of the phase pattern at every 12 ms (data from Participant #3). Time proceeds from left to right, from top to bottom. Each map shows a phase pattern of alpha activity on the scalp viewed from above (top is front, bottom is back), color-coded according to the color bar at the bottom. The letters a and b indicate the time points at which global phase-synchrony pattern shows changes. Phase values measured at 169 points on the scalp were linearly interpolated to obtain spatially continuous images

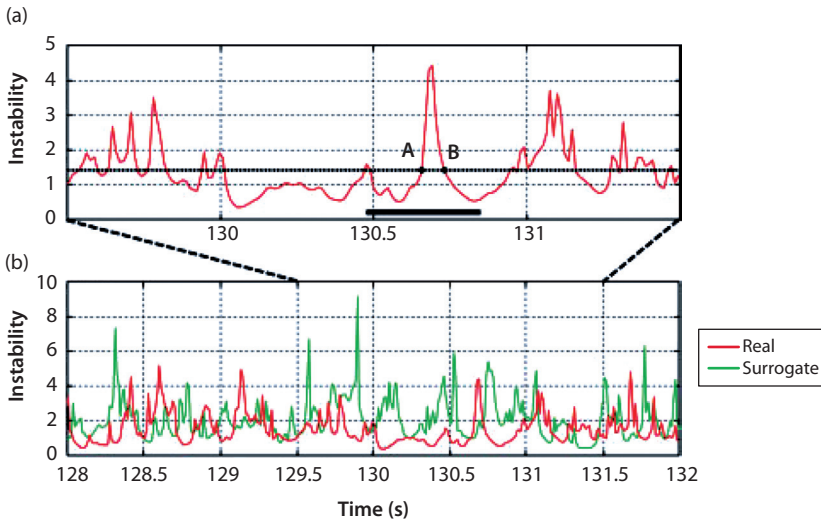


Figure 8. From Ito et al. (2007). (a) Time series of the instantaneous instability index I of a relative-phase pattern. An arbitrary threshold used to determine laminar intervals is shown with a dotted horizontal line. The wave patterns before *a* and after *b* reflect laminar periods. (b) Time series of I for the real and typical surrogate data shown in red and green lines, respectively

Ito et al. (2007) observed a dynamic behavior characteristic of chaotic itinerancy (Kaneko & Tsuda 2001; Tsuda 2001). Chaotic itinerancy occurs when fragile attractors are corrupted by noise. Unlike stable attractors, fragile attractors are manifolds on which not all neighbouring trajectories converge; in other words there are always points near the attractor, from which it is possible to escape. As a result, infinitesimal noise can switch the system from an approaching to an escaping trajectory. Noise-corrupted attractors are still preferably visited, as overall the system is attracted to them. The system will dwell in one of its neighborhoods for some time, until noise enables it to escape. Once away from the attractor, the system will show some irregular behavior, and eventually become attracted again.

Gong, Nikolaev & van Leeuwen (2007) showed that this behavior could be attributed to a critical threshold phenomenon: dynamic synchronization with $1/f$ characteristics in the durations of synchrony occurs in a system of coupled oscillators (Kuramoto 1975), when the coupling strength between the oscillators is at the critical threshold of mutual entrainment.

As previously mentioned, the signature $1/f$ behavior also occurs in alpha amplitude (Linkenkaer-Hansen et al. 2001). How are both related? Alpha amplitude modulation may be a gating mechanism. Inhibition operates in a

topologically more restricted region when the peaks have a higher frequency (i.e. shorter peak to peak intervals). As argued, this has its influence on the degree of context-sensitivity with which information is being processed *within* cortical regions. The intervals of synchronized activity may reflect the time it takes to communicate the information *between* regions, as synchronization facilitates communication between neural assemblies (Livanov 1977; von der Malsburg 1985). It helps to keep the temporal “windows of communication” concurrently open for input and output (Fries 2005).

The wave patterns that appear during the laminar periods might be considered the basis for the unity of experience. The phenomenal counterpart to the duration of a laminar period would be the psychological present (Stroud 1967). This simple equivalence statement has a particularly strong consequence: people fade out of consciousness during the brief periods of instability between laminar processes. This prediction should be taken with a few grains of salt. First, there is no critical threshold for laminar intervals in Figure 7, and therefore, by consequence neither is a critical threshold for consciousness implied here. One could, accordingly, be more or less conscious. One could simultaneously be aware in some modality and unaware in another – think of you suddenly realizing the awful smell while engaged in studying the color spectrum of Limburg cheese. Nevertheless, there may be some residual truth in the prediction: Consider that perceptual switching is experienced as instantaneous whereas the actual processes last 0.5–1.0 s (Ito, Nikolaev, Luman et al. 2003). During the switching, the system is on a transient between two laminar intervals (Nakatani & van Leeuwen 2005, 2006), so there is no persistent synchrony and hence no experienced duration.

Perhaps, the critical threshold will be discovered in the neural dynamics underlying consciousness. Indeed, if the dynamics is governed by fragile attractors, a system on an approaching trajectory would be conscious and one escaping from the attractor would not.

5. What happens to the laminar intervals when observers do engage in a task?

So far, we have been looking at the resting EEG. Once observers engage in a task, the large-scale patterns of alpha synchronization tend to disappear. Instead, we detect gamma or beta activity that accompanies the enhancement of a focus of mental activity in dedicated brain regions. Alpha activity plays a complementary role in other regions: it appears wherever objects or features need to be specifically

ignored or selected against (Foxe & Snyder 2011). Accordingly, in the Necker cube, switching was invariably preceded by transient patterns of synchrony in the gamma band (Nakatani & van Leeuwen 2006). These occur in parietal areas relevant to the emergence of a new perceptual organization (Nakatani, Orlandi & van Leeuwen 2011). Switching sometimes, but not always, was accompanied by transient alpha band activity in occipital areas. This activity may be related to the process of suppressing newly incoming sensory information to enable reading out information from the primary visual cortex (Shevelev, Kosteliantz, Kamenkovich et al., 2001).

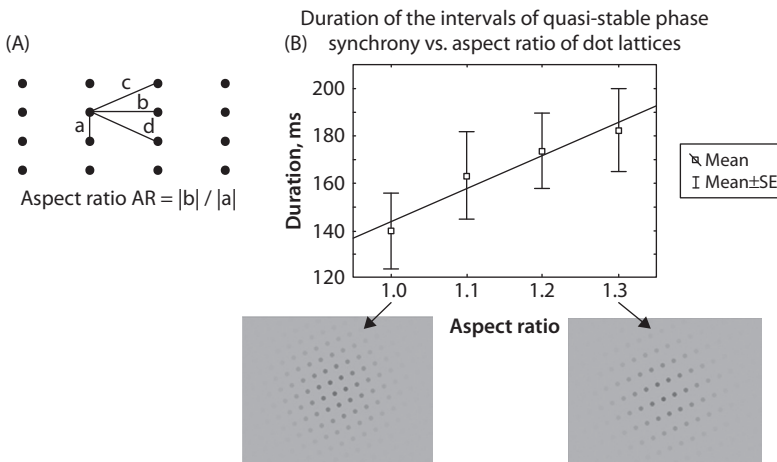


Figure 9. (adapted from Nikolaev et al., 2010). A. Grouping preference (columns vs. rows) as a function of Aspect Ratio (AR), based on the Gestalt principle of proximity. B. The mean ($N = 8$) duration of intervals of phase synchrony derived from evoked EEG in the parieto-occipital scalp areas as a function of AR. Dot lattice stimuli corresponding to $AR = 1.0$ and $AR = 1.3$ are shown under the graph

As the characteristic whole-head alpha patterns disappear, preferred durations will emerge in the intervals of synchrony. It then becomes meaningful to consider, for instance, effects on duration averages. We adopted a psychophysical paradigm in which the perceived grouping of dot lattices is quantitatively determined as a function of a stimulus parameter: proximity (Kubovy et al. 1998). Proximity determines perceived grouping through a simple relationship called Aspect Ratio (AR) see Figure 9A. The larger AR the stronger is the preference for grouping according to the smallest distance (a); the more AR approaches 1, the more ambiguous is the perception of the lattice. Note that ambiguity equals uncertainty, or lack of information.

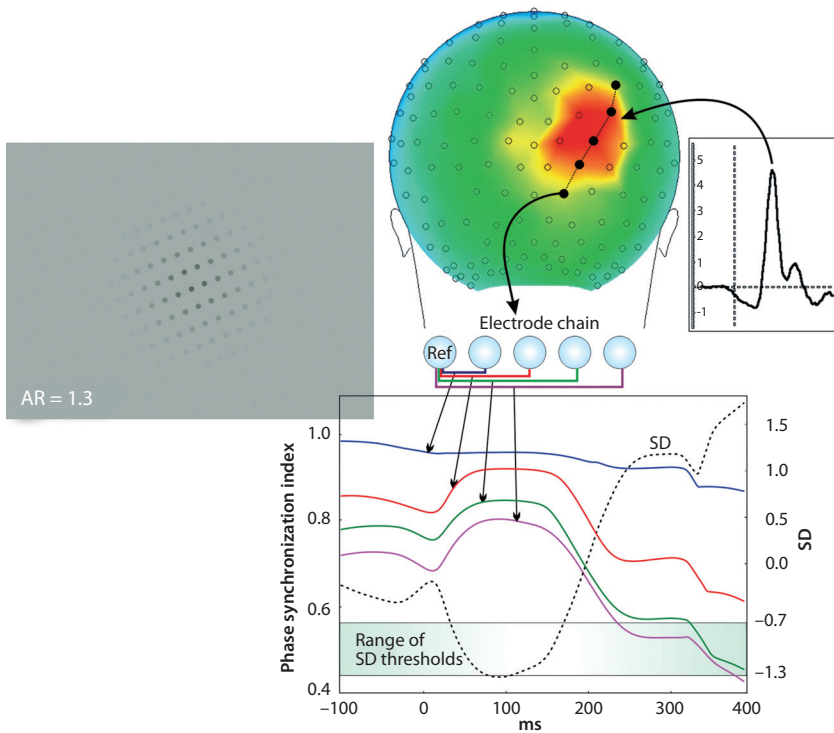


Figure 10. (adapted from Nikolaev et al., 2010). Following presentation of a dot-lattice (left panel) to an observer, evoked activity appears (rightmost panel). For this evoked activity we consider the pair-wise dynamic synchronization of an array of electrodes with a reference (Ref) depicted in the lowermost panel. Due to volume conductivity, the synchronization decreases, on average, with distance. Sometimes, however, the whole array takes on more uniform synchronization values, regardless of distance. This is expressed in their standard deviations (SD). We chose a threshold and define as a coherence interval the time the SD remains below the threshold

We then determined which evoked component was sensitive to AR (Nikolaev et al. 2008b). At the scalp location of that component, we measured the durations of synchronized intervals in relation to the aspect ratio of the dot lattice (Figure 8A). We focused on the beta rhythm, which has been associated to visual attention and object related processing (Gross, Schmitz, Schnitzler et al. 2004; Tallon-Baudry, Bertrand, Fischer 2001; von Stein, Rappelsberger, Sarnthein et al. 1999, see also Wrobel 2000), and is of particular importance for our next example. We found a simple, linear relation of aspect ratio with coherence interval duration (Figures 9B; 10). Aspect ratio lawfully determines the ambiguity of the stimulus: the greater the aspect ratio, the less ambiguous the stimulus. Ambiguity is the

absence of information. Therefore, by definition, the less ambiguous a stimulus, the more information it contains. Thus, the more information in the stimulus, the longer the interval of synchrony the stimulus evokes. The steepness of the slope in individuals was found to be related to how sensitive they are to aspect ratio. Interestingly, the intervals of synchrony show extreme-value distributions, corresponding to the idea that the propagation of information takes place in parallel across multiple channels, and that the length of the interval is determined by the slowest of these processes (cf. Pöppel 1970). Intervals of synchronized activity, thus, reflect the time needed for intracortical communication of visual information. We concluded that the intervals of synchronized activity directly reflect the amount of stimulus information, as promulgated by the visual system to the rest of the brain.

Even though the episodes of synchrony occurring in “rest” alpha and those in the beta range are quite different, we might speculate that they both have similar information-processing functions, as well as being the vehicles of consciousness. Despite the differences in signal (alpha vs beta), the results from Smit and van Leeuwen’s (1994) Necker cube and Nikolaev et al.’s dot lattice studies point consistently to the same principle (van Leeuwen 1998): slower peak alpha and longer beta coherence intervals both indicate that more information is being integrated in the neurons, circuits, and brain regions involved in concurrent processing.

6. Conclusion

We have seen that mind wandering is not necessarily a process of distraction; minds engage in wandering to raise the quality of their percept. Greater consistency within a figure, and between the figure and its context, are the result. Artists, designers, and perceivers of art make use of mind-wandering to create, or detect, surplus structure. We found that wandering shows long-term regularities in its dynamics, which could be detected both in experience (the Necker cube) and in brain activity (rest alpha). It would therefore appear that “resting” is not the most appropriate term, when the mind is actively wandering.

Acknowledgement

Authors would like to thank Dr. Andrey R. Nikolaev and the editors of this volume for valuable comments on an early version of this paper.

References

- Angelakis, E., Lubar, J.F., Stathopoulou, S. & Kounios, J. (2004). Peak alpha frequency: An electroencephalographic measure of cognitive preparedness. *Clinical Neurophysiology*, 15 (4), 887–897.
- Anokhin, A.P. & Vogel, F. (1996). EEG alpha rhythm frequency and intelligence in normal adults. *Intelligence*, 23, 1–14.
- Barlow, H.B., Földiák, P. (1989). Adaptation and decorrelation in the cortex. In Durbin, R., Miall, C., Mitchison, C. (Eds.), *The Computing Neuron* (pp. 54–72). Addison-Wesley, Workingham.
- Bhattacharya, J., Petsche, H., Feldmann, U. & Rescher, B. (2001). EEG gamma-band phase synchronization between posterior and frontal cortex during mental rotation in humans. *Neuroscience Letters*, 311, 29–31.
- Bassingthwaight, J.B., Liebovitch, L.S. & West, B.J. (1994). *Fractal physiology*. New York, NY: Oxford University Press.
- Braboszcz, C., & Delorme, A. (2011). Lost in thoughts: Neural markers of low alertness during mind wandering. *NeuroImage*, 54, 3040–3047.
- Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science*, 7, 404–416.
- Boenke, L.T., Ohl, F., Nikolaev, A.R., Lachmann, T. & van Leeuwen, C. (2009). Stroop and Garner interference dissociated in the time course of perception, an event-related potentials study. *NeuroImage*, 45, 1272–1288.
- Borsellino, A., De Marco, A., Allazetta, A., Rinesi, S. & Bartolini, B. (1972). Reversal time distribution in the perception of visual ambiguous stimuli. *Kybernetik*, 10, 139–144.
- Crowne, D.P. (1983). The frontal eye field and attention. *Psychological Bulletin*, 93, 232–260.
- Breakspear, M., Williams, L.M. & Stam, C.J. (2004). A novel method for the topographic analysis of neural activity reveals formation and dissolution of ‘dynamic cell assemblies’. *Journal of Computational Neuroscience*, 16, 49–68.
- Ditzinger, T. & Haken, H. (1989). Oscillations in the perception of ambiguous patterns. *Biological Cybernetics*, 61, 279–287.
- Ditzinger, T. & Haken, H. (1990). The impact of fluctuations on the recognition of ambiguous patterns. *Biological Cybernetics*, 63, 453–456.
- Doppelmayr, M., Klimesch, W., Stadler, W., Pöllhuber, D. & Heine, C. (2002). Alpha power and intelligence. *Intelligence*, 30, 289–302.
- Einhäuser, W., Martin, K.A.C. & König, P. (2004). Are switches in perception of the Necker cube related to eye position? *European Journal of Neuroscience*, 20, 2811–2818.
- Fell, J., Klaver, P., Lehnertz, K., Grunwald, T., Schaller, C., Elger, C.E. & Fernández, G. (2001). Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling. *Nature Neuroscience*, 4, 1259–1264.
- Foxe, J.J. & Snyder, A.C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers in Psychology*, Art 154.
- Freeman, W.J. & van Dijk, B.W. (1987). Spatial patterns of visual cortical fast EEG during conditioned reflex in a rhesus monkey. *Brain Research*, 422(2), 267–276.
- Freeman, W.J. & Rogers, L.J. (2002). Fine temporal resolution of analytic phase reveals episodic synchronization by state transitions in gamma EEGs. *Journal of Neurophysiology*, 87, 937–945.

- Fries P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Science*, 9, 474–480.
- Gao, J.B., Billock, V.A., Merk, I., Tung, W.W., White, K.D., Harris, J.G. & Roychowdhury, V.P. (2006). Inertia and memory in ambiguous visual perception, *Cognitive Processing*, 7, 105–112.
- Garner, W.R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.
- Gaser, T., Von Lucadou-Muller, I., Verleger, R. & Bacher, P. (1983). Correlating EEG and IQ: A new look at an old problem using computerized EEG parameters. *Electroencephalography and Clinical Neurophysiology*, 55, 493–504.
- Gepshtein, S. & Kubovy, M. (2007). The lawful perception of apparent motion. *Journal of Vision*, 7, 1–15.
- Gilden, D.L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, 8, 296–301.
- Gilden, D.L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, 108, 33–56.
- Gilden, D.L., Thornton, T. & Mallon, M.W. (1995). 1/f noise in human cognition. *Science*, 267, 1837–1839.
- Gong, P., Nikolaev, A.R. & van Leeuwen, C. (2003). Scale-invariant fluctuations of the dynamical synchronization in human brain electrical activity. *Neuroscience Letters*, 336, 33–36.
- Gong, P., Nikolaev, A.R. & van Leeuwen, C. (2007). Dynamics of collective phase synchronization in human electrocortical activity. *Physical Review E*, 76, art. no. 011904.
- Green, D.M. & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shapiro, K., Hommel, B. & Schnitzler, A. (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proceedings of the National Academy of Sciences, USA*, 101, 13050–13055.
- Gruber, T., Müller, Keil, A. & Elbert, T. (1999). Selective visual-spatial attention alters induced gamma band responses in the human EEG. *Clinical Neurophysiology*, 110, 2074–2085.
- Healy, J. & Slater, D. (1994). Global color constancy: Recognition of objects by use of illumination-invariant properties of color distributions. *Journal of the Optical Society of America A*, 11, 3003–3010.
- Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79, 2554–2558.
- Inouye, T., Shinosaki, K., Toi, S., Matsumoto, Y. & Hosaka, N. (1995). Potential flow of a activity in the human electroencephalogram. *Neuroscience Letters*, 187, 29–32.
- Ito, J., Nikolaev, A.R., Luman, M., Aukes, M.F., Nakatani, C. & van Leeuwen, C. (2003). Perceptual switching, eye-movements, and the bus-paradox. *Perception*, 32, 681–698.
- Ito, J., Nikolaev, A.R. & van Leeuwen, C. (2005). Spatial and temporal structure of phase synchronization of spontaneous EEG alpha activity. *Biological Cybernetics*, 92, 54–60.
- Ito, J., Nikolaev, A.R. & van Leeuwen, C. (2007). Dynamics of spontaneous transitions between global brain states. *Human Brain Mapping*, 28, 904–913.
- Jaarsveld, S. & van Leeuwen, C. (2005). Sketches from a design process: Creative cognition inferred from intermediate products. *Cognitive Science*, 95, 79–107.
- Jaušovec, N. & Jaušovec, K. (2000). Correlations between ERP parameters and intelligence: A reconsideration. *Biological Psychology*, 55, 137–154.
- Kaneko, K. & Tsuda, I. (2001). *Complex systems: Chaos and beyond* – A constructive approach with applications in life sciences. Berlin: Springer Verlag.

- Kelly, A., Heathcote, A., Heath, R. & Longstaff, M. (2001). Response time dynamics: Evidence for linear and low-dimensional nonlinear structure in human choice sequences. *Quarterly Journal of Experimental Psychology*, 54A, 805–840.
- Kim, Y.-J., Grabowecky, M. & Suzuki, S. (2006). Stochastic resonance in binocular rivalry. *Vision Research*, 46, 392–406.
- Köhler, W. & Wallach, H. (1944). Figural after-effects: An investigation of visual processes. *Proceedings of the American Philosophical Society*, 88, 269–357.
- Kubovy, M., Holcombe, A.O. & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology*, 35, 71–98.
- Kuramoto K., (1975). Self-entrainment of a population of coupled non-linear oscillators. *Lecture Notes in Physics*, 39, 420–422.
- Laufs, H., Holt, J.L. et al. (2006). Where the BOLD signal goes when alpha EEG leaves. *NeuroImage*, 31(4), 1408–1418.
- Lachmann, T. & van Leeuwen, C. (2008). Differentiation in holistic processing in letter versus non-letter shape recognition. *Acta Psychologica*, 129, 121–129.
- Lavie, N. & Cox, S. (1997). On the efficiency of selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, 8, 395.
- Lehmann, D., Ozaki, H., & Pal, I. (1987). EEG alpha map series: Brain micro-states by space-oriented adaptive segmentation. *Electroencephalography and Clinical Neurophysiology*, 67, 271–288.
- Levelt, W.J.M. (1967). Note on the distribution of dominance times in binocular rivalry. *British Journal of Psychology*, 58, 143–145.
- Linkenkaer-Hansen, K., Nikouline, V.V., Palva, S. Ilmoniemi, R.J. & Palva, J.M. (2004). Prestimulus oscillations enhance psychophysical performance in humans. *Journal of Neuroscience*, 24(45), 10186–10190.
- Linkenkaer-Hansen, K., Nikouline, V.V., Palva, J.M. & Ilmoniemi, R.J. (2001). Long-range temporal correlations and scaling behavior in human brain oscillations. *Journal of Neuroscience*, 21, 1370–1377.
- Livanov, M.N. (1977). *Spatial organization of cerebral processes*. New York: John Wiley & Sons.
- Lopes da Silva, F.H. (1991). Neural mechanisms underlying brain waves: from neural membranes to networks. *Electroencephalography and Clinical Neurophysiology*, 79, 81–93.
- Lucassen, M.P. & Walraven, J. (1993). Quantifying color constancy: Evidence for nonlinear processing of cone-specific contrast. *Vision Research*, 33, 739–757.
- Mandelbrot, B.B. & Wallis, J.R. (1969). Some long-run properties of geophysical records. *Water Resources Research*, 5, 321–340.
- Maturana, H.R. & Varela, F.J. (1980). *Autopoiesis and cognition: The realization of the living*. Boston studies in the philosophy of science, vol. 42. Dordrecht, NL.: D. Reidel Publ.
- Mantini, D., Perrucci, M.G. et al. (2007). Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences of the USA*, 104, 13170–13175.
- Mathewson, K.E., Lleras, A., Beck, D.M., Fabiani, M., Ro, T. & Gratton, G. (2011). Pulsed out of awareness: EEG alpha oscillations represent a pulsed-inhibition of ongoing cortical processing. *Frontiers in Psychology*, 2, Article 99.
- Nakatani, C., Ito, J., Nikolaev, A.R., Gong, P.L. & van Leeuwen, C. (2005). Phase synchronization analysis of EEG during attentional blink. *Journal of Cognitive Neuroscience*, 17, 1969–1979.
- Nakatani, H. & van Leeuwen, C. (2005). Individual differences in perceptual switching rates; The role of occipital alpha and frontal theta band activity. *Biological Cybernetics*, 93, 343–354.

- Nakatani, H. & van Leeuwen, C. (2006). Transient synchrony of distant brain areas and perceptual switching. *Biological Cybernetics*, 94, 445–457.
- Nakatani, H., Orlandi, N., & van Leeuwen, C. (2011). Precisely timed oculomotor and parietal EEG activity in perceptual switching. *Cognitive Neurodynamics*, 5, 399–409.
- Necker, L.A. (1832). Observation on some remarkable optical phaenomena seen in Switzerland; and on an optical phaenomenon, which occurs on viewing a figure of a crystal or geometrical solid. *London and Edinburgh Philosophical Magazine and Journal of Science*, 1, 329–337.
- Nikolaev, A.R., Gepshtein, S., Kubovy, M. & van Leeuwen, C. (2008). Dissociation of early evoked cortical activity in perceptual grouping. *Experimental Brain Research*, 186(1), 107–122.
- Nikolaev, A.R., Gepshtein, S., Gong, P. & van Leeuwen, C. (2010). Duration of coherence intervals in electrical brain activity in perceptual organization. *Cerebral Cortex*, 20, 365–382.
- Nunez, P.L. (1974). Wavelike properties of the alpha rhythm. *IEEE Transactions on Biomedical Engineering*, 21, 473–482.
- Osaka, M. (1984). Peak alpha frequency of EEG during a mental task: task difficulty and hemispheric differences. *Psychophysiology*, 21(1), 101–105.
- Ozaki, H. & Suzuki, H. (1987). Transverse relationships of the alpha rhythm on the scalp. *Electroencephalography and Clinical Neurophysiology*, 66, 191–195.
- Palva, S. & Palva J.M. (2007). New vistas for alpha-frequency band oscillations. *Trends in Neurosciences*, 30(4), 150–158.
- Peng, C.-K., Havlin, S., Stanley, H.E. & Goldberger, A.L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*, 5(1), 82–87.
- Peterson, M.A. & Gibson, B.S. (1991). Directing spatial attention within an object: Altering the functional equivalence of shape descriptions. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 170–182.
- Peterson, M.A. & Hochberg, J. (1983). Opposed-set measurement procedure: A quantitative analysis of the role of local cues and intention in form perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 183–193.
- Petsche, H. & Sterc, J. (1968). The significance of the cortex for the travelling phenomenon of brain waves. *Electroencephalography and Clinical Neurophysiology*, 25, 11–22.
- Pöppel, E. (1970). Excitability cycles in central intermittency. *Psychologische Forschung*, 34, 1–9.
- Posthuma, D., Neale, M.C., Boomsma, D.I. & de Geus, E.J.C. (2001). Are smarter brains running faster? Heritability of alpha peak frequency, IQ, and their interrelation. *Behavior Genetics*, 31, 567–579.
- Rock, I. & Gutman, D. (1981). The effect of inattention on form perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 275–285.
- Sanchez-Vives, M.V., Nowak, L.G. & McCormick, D.A. (2000). Membrane mechanisms underlying contrast adaptation in cat area 17 *in vivo*. *Journal of Neuroscience*, 20, 4267–4285.
- Sarnthein, J., Petsche, H., Rappelsberger, P., Shaw, G.L. & von Stein, A. (1998). Synchronization between prefrontal and posterior association cortex during human working memory. *Proceedings of the National Academy of Sciences, USA*, 95, 7092–7096.
- Schack, B., Weiss, S. & Rappelsberger, P. (2003). Cerebral information transfer during word processing: Where and when does it occur and how fast is it? *Human Brain Mapping*, 19, 18–36.
- Schooler, J.W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Science*, 6, 339–344.
- Silberstein, R.B., Danieli, F. & Nunez, P.L. (2003). Fronto-parietal evoked potential synchronization is increased during mental rotation. *Neuroreport*, 14, 67–71.

- Shaw, J.C. (2004). *The brain's alpha rhythms and the mind*. Amsterdam, NL: Elsevier Science.
- Shaw, J.C. & McLachlan, K.R. (1968). The association between alpha rhythm propagation time and level of arousal. *Psychophysiology*, 4, 307–310.
- Shevelev, I.A., Kostelianetz, N.B., Kamenkovich, V.M. & Sharaev, G.A. (2001). EEG alpha-wave in the visual cortex: Check of the hypothesis of the scanning process. *International Journal of Psychophysiology*, 11, 195–201.
- Skarda, C.A. & Freeman, W.J. (1987). How brains make chaos to make sense of the world. *Behavioral and Brain Sciences*, 10, 161–195.
- Smallwood, J. & Schooler, J.L. (2006). The restless mind. *Psychological Bulletin*, 132, 946–958.
- Smit, D.J.A., Heslenfeld, D., Kenemans, L., Kohl, K. & van Leeuwen, C. (1994). Global organization influences the perceived orientation of the Necker cube. Unpublished manuscript.
- Smit, C.M., Wright, M.J., Hansell, N.K., Geffen, G.M. & Martin, N.G. (2006). Genetic variation of individual alpha frequency (IAF) and alpha power in a large adolescent twin sample. *International Journal of Psychophysiology*, 61, 235–243.
- Sokolov, A., Lutzenberger, W., Pavlova, M., Preissl, H., Braun, Ch. & Birbaumer, N. (1991). Gamma-band MEG activity to coherent motion depends on task-driven attention. *Neuroreport*, 10, 1997–2000.
- Stroud, J.M. (1967). The fine structure of psychological time. *Annals of the NY Academy of Science*, 138, 623–631.
- Tallon-Baudry, C., Bertrand, O. & Fischer, C. (2001). Hebbian reverberating oscillatory synchrony during memory rehearsal in the human brain. *Journal of Neuroscience*, 21, 1–5.
- Taylor, M.M. & Aldridge, K.D. (1974). Stochastic processes in reversing figure perception. *Perception & Psychophysics*, 16, 9–27.
- Thatcher, R.W., Krause, P. & Hrybyk, M. (1986). Corticocortical associations and EEG coherence – a 2-compartmental model. *Electroencephalography and Clinical Neurophysiology*, 64, 123–143.
- Thier, P., Haarmeier, T., Chakraborty, S., Lindner, A. & Tikhonov, A. (2001). Cortical substrates of perceptual stability during eye movements. *NeuroImage*, 14, 33–39.
- Tsuda, I. (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and Brain Sciences*, 24, 793–810.
- van Dam, L.C.J. & van Ee, R. (2006). The role of saccades in exerting voluntary control in perceptual and binocular rivalry. *Vision Research*, 46, 787–799.
- Van De Ville, D., Britz, J. & Michel, C.h.M. (2010). EEG microstate sequences in healthy humans at rest reveal scale-free dynamics. *Proceedings of the National Academy of Sciences, USA*, 107, 18179–18184.
- van Ee, R. (2009). Stochastic variations in sensory awareness are driven by noisy neuronal adaptation: Evidence from serial correlations in perceptual bistability. *Journal of the Optical Society of America A*, 26, 2612–2622.
- van Leeuwen, C. (1998). Visual perception at the edge of chaos. In J.S. Jordan (Ed.), *Systems Theories and a-priori aspects of perception* (pp. 289–314). Amsterdam, NL: Elsevier.
- van Leeuwen, C. (2007). What needs to emerge to make you conscious? *Journal of Consciousness Studies*, 14, 115–136.
- van Leeuwen, C. & Lachmann, T. (2004). Negative and positive congruence effects in letters and shape. *Perception & Psychophysics*, 66, 908–925.
- van Leeuwen, C., Verstijnen, I.M. & Hekkert, P. (1999). Common unconscious dynamics underlie uncommon conscious effect: A case study in the iterative nature of perception and

- creation. In J.S. Jordan (Ed.), *Modeling consciousness across the disciplines* (pp. 179–218). Lanham, MD: University Press of America.
- Van Orden, G.C., Holden, J.G. & Turvey, M.T. (2003). Selforganization of cognitive performance. *Journal of Experimental Psychology: General*, 132, 331–350.
- Varela, F., Lachaux, J.-P., Rodriguez, E. & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2, 229–239.
- Verstijnen, I.M., van Leeuwen, C., Goldschmidt, G., Hamel, R. & Hennessey, J.M. (1998a). Creative discovery in imagery and perception: Combining is relatively easy, restructuring takes a sketch. *Acta Psychologica*, 99, 177–200.
- Verstijnen, I.M., van Leeuwen, C., Goldschmidt, G., Hamel, R. & Hennessey, J.M. (1998b). Sketching and Creative Discovery. *Design Studies*, 19, 519–546.
- Verstijnen, I.M., van Leeuwen, C., Hamel, R. & Hennessey, J. (2000). What imagery can't do and why sketching might help. *Empirical Studies of the Arts*, 18, 167–182.
- von der Malsburg, C. (1985). Nervous structures with dynamical links. *Berichte der Bunsen-Gesellschaft für Physikalische Chemie*, 89, 703–710.
- von Stein, A., Rappelsberger, P., Sarnthein, J. & Petsche, H. (1999). Synchronization between temporal and parietal cortex during multimodal object processing in man, *Cerebral Cortex*, 9, 137–150.
- Wagenmakers, E.J., Grünwald, P. & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50, 149–166.
- Witkin, H.A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality*, 19, 1–15.
- Wrobel, A. (2000). Beta activity: A carrier for visual attention. *Acta Neurobiologica Experimentalis*, 60, 247–260.

Fuzzy consciousness

Stephanie Huetten & Michael Spivey

University of California, Merced

This chapter explores the consequences of treating consciousness as a fuzzy dynamical system. A fuzzy dynamical system is one in which labeled concepts and percepts are altered as a function of context and conditions, and these changes occur continuously in time. We offer speculations on the groundwork for a consciousness state space in which the sets of trajectories over time form tube-like structures called cylinder sets. Consciousness is a trajectory within this structure, and it passes by or through those fuzzy concepts. When the trajectory of a mental event travels close to a particular concept or percept, one experiences awareness of that concept or percept. As these constrained pathways in mental state space become more heavily traveled, they develop increased density (or attraction strength) in their central threads, and more and more nearby trajectories get captured by that cylinder. At the same time, these tubes slowly gravitate toward short cuts in the state space over the lifespan, thus gradually straightening out and skipping past intermediating concepts that used to get visited as part of the sequence. The fringes of these concepts become a part of conscious experience, and for everyday coping, do not need to be recruited for explicit awareness. The formation and streamlining of cylinder sets over the course of learning may have the paradoxical result of producing an increase in tacit conscious experience (of “being in time”) and a decrease in explicit awareness (of this or that labeled concept).

1. Introduction

The building blocks of a physical science are based in equations with time as a variable. This variable can be a scalar, such as the duration of an event, or a vector describing changes in distance and direction of an object over time (velocity). From equations we can predict where objects will fall after being thrown into the air, how fast an airplane must accelerate to takeoff within a given time interval, or even when someone is about to gain insight on how to better perform a task (Stephen, Boncodd, Magnuson & Dixon 2009). Time is a useful concept to have, because of its predictive power. From another view, time is a useful concept for equations and descriptions of physical phenomena, but could be epiphenomenal,

stemming from the need to describe change (Barbour 2001). Consciousness is typically addressed from a *why* framework: why it exists, what it is made of, if we're the only species that has it. The *why* framework is purely philosophical, where one makes a logical argument to address a question that cannot be operationalized in the laboratory (e.g. we cannot run an experiment on *why* consciousness exists). This branch of thinking is easily identifiable because it usually begins with a "why" question and is interested in qualia, verbal reports, and definitions of subcategories of consciousness and awareness.

More than the method of how one answers these questions, it is arguably the question itself that leads us to using various methods, which in turn leads us to different conclusions. Thus, a major goal of this introduction is to encourage a different kind of question that can be asked about consciousness. We will call this the *what* framework. Just as a parabola describes the path of a falling object on Earth, so can we describe a path through a mental space. In describing movement along any path, the obvious requisite is a time variable. The description of physical motion says nothing about the mechanism of gravity, why it exists, or where it came from. A description of consciousness that includes time may also need to concede to not immediately answering these questions. This is not to say that a *why* framework is unimportant: rather it will come organically and naturally from the equations. Just as our understanding of quantum physics is developing and changing based upon the mathematics of the field, so do our descriptions of the mind need to build a foundation for our theory.

The most difficult part of transitioning to a descriptive, predictive, controllable model of researching consciousness will be abandoning the idea that each outcome can be traced to one causal force, or a simple additive combination of causal forces. In traditional psychological research, one uses a stimulus and carefully manipulates one variable, resulting in some behavioral difference. In recent years, evidence of interactions among variables has encouraged testing multiple variables at once, and measuring behavior from a sensory modality outside of the one being stimulated (see Spivey 2007, Chapter 3). Through this kind of paradigm we have discovered evidence of a cognitive system that is underlyingly interactive (Stephen & Mirman 2010; Spence & Driver 2004), exhibits characteristics of complexity and interactivity (Kello, Anderson, Holden & Van Orden 2008; Kello & Van Orden 2009), and has many top-down constraints on low-level processes (Motter 1993; Spivey & Spirn 2000; Gandhi, Heeger & Boynton 1999; Ito & Gilbert 1999; Lamme & Roelfsema 2000). All of these principles have something in common: they all imply that everyday conscious experience is noisy and interactive. Most of one's time is spent moving through a mental state space (whose axes are defined by neural, kinematic and environmental parameters) containing concepts, percepts, and emotions, never quite fully reaching any one pure concept in a context-free

manner. This means that at any given time, consciousness is defined by the many thousands of environmental inputs in a natural environment, the constraints of many billions of neurons with previous learning experience, and a body equipped with high-dimensional sensation and movement parameters. All of these variables collude to form what is experienced as consciousness.

And so by abandoning linear-additive causality, one does not need to describe a trillion-dimensional state space by listing causal forces for weeks at a time. For instance, pretend Tom is moving toward fuzzy conscious state A. Along the way, he passes through some tempered anger, while processing sensory information from his messy apartment. He has been frustrated with himself for not being neater so he can invite his girlfriend over for dinner, and his boss yelled at him earlier that day. On top of this, he constantly day dreams of being a rock star and living on easy street, while struggling with self-esteem issues developed during his childhood. Finally, we get near the fuzzy conscious state A, which produces something unexpected: Tom begins to laugh. Each factor contributed to Tom arriving near fuzzy conscious state A, but none of them can individually explain why Tom is suddenly laughing. It could be that he suddenly remembered a joke; it could be that everything in his life suddenly doesn't seem quite so serious and he is making fun of himself in a playful way. Whatever the case may be, every single movement a person makes through the state space of consciousness is a function of thousands of variables at multiple timescales (Van Orden & Holden 2002). Every millisecond as one traverses through life is part of the building, and becomes a trace, or an integral part of the experience of consciousness. Attempting to identify individual causal forces for consciousness is almost insulting: the complexity of human experience will not be explained by one moment or one variable. When done right, the characterization of causal forces that led to the emergence of a particular state of consciousness will be a description reduced in dimensionality such that visualization and analysis are possible, but will still contain much of the important information from the original high-dimensional space.

What this experience *is like* to someone, we argue, is substantially a product of linguistic report, which is only an approximation of what the experience may actually be. A verbal report is not what the experience is; it is a report of what the experience *is like*. Many studies have demonstrated that what people report their mental processes are like is unreliable (Nisbett & Wilson 1977), inaccurate (Wells 1984), and can even interfere with cognitive processing (Schooler, Ohlsson & Brooks 1993). Unfortunately, the only way to communicate experience is via some form of language. Thus, when using language as a dependent measure, we get a snapshot of how someone reports an inner examination, not the experience itself. Further, it would be hard to hang onto an accurate description of how a conscious state evolves on a millisecond timescale, much

less be able to convey this linguistically. Any verbal report is best described as a very coarse lump sum.

This chapter contains some aspects of how consciousness may emerge during the mind's traversal through a dynamic state space of thoughts, actions and concepts. Many regions of this state space generate very little experience of consciousness. However, other parts of this state space are associated with a tacit conscious experience of "being in time," without necessarily self-reflecting on the meanings and labels of the percepts and concepts being entertained. Other parts of the state space are also associated with awareness of the meanings and labels being visited. The goal here is to provide a framework for an intuitive characterization of a conscious system on a millisecond timescale, as well as the timescales of days and weeks, and to examine how *consciousness* (i.e. being in time) and *awareness* (i.e. of percepts and concepts) interact and develop over a lifetime. Special consideration will also be given to cultural factors and other practices which are hypothesized to affect human consciousness.

2. Cylinder sets

An infant is born with a set of physical and genetic constraints on its brain and body. Certain areas of the brain such as the visual cortex will be predisposed to developing a particular kind of connectivity pattern. Eye color is determined by genetic factors. A set of predispositions will already be in place for things such as athletic ability, though gene expression can be substantially altered by environmental factors (Choi & Kim 2007; Jaenisch & Bird 2003; Rampon et al. 2000). The interplay of genes and environment are roughly what sandwich the mechanisms that drive the formation of constraints (Elman et al. 1996, Chapters 1 and 6). They define what can be experienced, and in turn this is ultimately what we have to work with when forming a cohesive theory of consciousness: the brain, the body, and the environment.

To present a cursory overview of how, in this framework, a child develops awareness and consciousness over the course of a lifetime, we'll go through a very rough chronological outline. Early in life, meaningless neural connections die off, as meaningful connections are maintained and this continues throughout childhood (Huttenlocher 1990). This is a process whereby the organism constrains the influence from its environment, i.e. putting what is sensed into meaningful categories and sharper percepts. For example, in the womb low-frequency sound patterns already allow the organism to begin distinguishing the mother's voice from those of others, and early in infancy phonetic features begin being categorized and color vision is developed. As the infant moves and thinks and discovers

new information, the information becomes more and more refined. If one imagines condensing the information down into a low-dimensional space, clusters in that space would correspond to different meanings. For example, human production of vowel sounds tend to form fuzzy partially-overlapping clusters as a function of fundamental frequency and the first two formants (peaks of energy) in the sound spectrum. When perceiving this information, the brain picks up on statistical patterns in the input (Saffran, Aslin & Newport 1996; Maye, Werker & Gerken 2002; Onnis, Waterfall & Edelman 2008). The particular frequencies that existed in the environment undergo some transformations when being internalized, but essential statistical structure is nonetheless preserved.

To illustrate this concept, Figure 1 shows some snapshots of how to envision the progression of this kind of clustering. Each point corresponds to a low-dimensional projection of a high-dimensional space. The clusters could correspond to a fuzzy concept or an action: their meaning is more in how this space is moved through, rather than its architecture. In the first panel, you can see that the state space is almost purely random, with only small hints of clusters. In the second panel, you can see that the state space is beginning to form some degree of structure, perhaps in the process of drifting toward an area of state space that would constitute a broad fuzzy category. Finally, the last panel depicts a macro level of what the structured organization of many thoughts (of various shapes) might look like in this state space, where many facets of cognition and perception are layered on one another, and a fractal structure of clusters within clusters has formed.

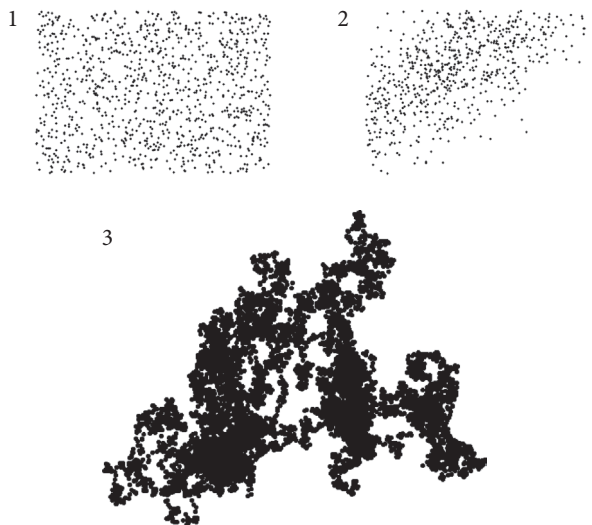


Figure 1. Concept formation over time

The pattern of points may correspond to any number of theories: the points could be a low-dimensional projection of activations from a high-dimensional episodic memory or connection strengths that embody the child's naïve "theories" about how the world works. For the formation of a theory of consciousness, the static structure or nature of concepts is not as important as how the mind moves through these concepts in time. Our definition of consciousness is then not anything about how *special* our feelings or thoughts are to us, nor does it have anything to do with explicit *awareness* of mental states and self-reflection. Rather consciousness is the ebb and flow between thoughts, feelings, external perturbations and internal fluctuations. It is not in itself an entity, but rather a phenomenon that results from moving through a high-dimensional biological network that is coupled with a dynamic environment. And so to characterize consciousness, we must describe how one travels from one thought to another.

Back to the infant network, once it has begun to learn about the world and perform transformations to the inputs it receives, small, connected trajectories begin to form. At first, it is not possible to move through the state space purposefully, because there are only very weak attractors barely pushing and pulling at the trajectory. However, as these fuzzy categories begin to condense and form a more cohesive structure, they begin to function as robust attractors and can string together a quite long and complex trajectory. This does predict that early in life, infants have very little consciousness (according to our definition), needing to have the flexibility of moving about rather freely in this space until well-traversed paths have been formed.

This then brings us to the cylinder sets, which we might envision as a natural road that goes from one popular location (e.g. point attractor) to another. In nature animals traveling through a field or forest trample down foliage, and paths are eventually formed. Similarly, movement through areas of mental space wears down a pathway. Here, we place emphasis on the portion of a cylinder set that lies *in between* the partitioned regions in state space that enjoy symbolic/conceptual labels (for further discussion, see beim Graben, Gerth & Vasishth 2008). Figure 2 is an example of what this end-goal of forming consciousness could be: a pathway that marks a common road that need not be directly through a concept, but rather at various distances to concepts and thoughts. The distance from the cylinder set to the concept's core of its attractor basin is roughly what we will define here as awareness. If the path travelled in mental state space goes right through center of the concept, then that concept is (at least briefly) the core object of awareness. However, if the path merely skips along the outskirts of an attractor basin, then that concept is only slightly a part of awareness. Thus, we can have many things *in the back of our mind* simply because the trajectory we are on is further from those thoughts, but still in that general area. William James called this a stream

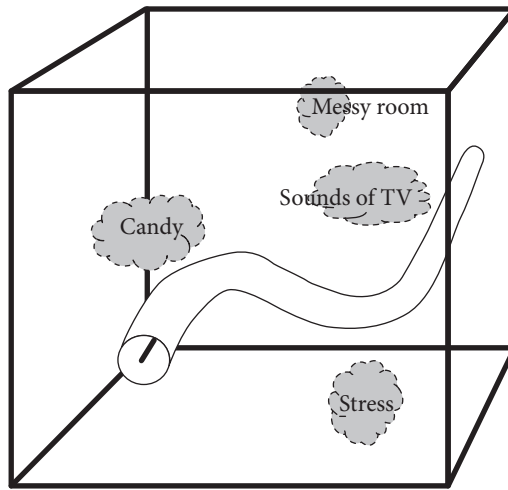


Figure 2. Cylinder set passing by fuzzy concepts

of thought, or stream of consciousness. Thomas Hobbes introduced the idea of a train of thought in the 1600's. The idea that consciousness is a stream, or taking place with contiguity in space and time, is not a new one.

Now we have cylinder sets of various well-formedness that tend to go in and around various thoughts. These well-traversed pathways manifest in action as well as in thought. As a simple example, consider walking. The pattern of behavior is cyclical: left step, right step, left step, right step. However, the execution of these motor commands exhibits long-range correlations, meaning the fluctuations in stride length exhibit recurring patterns at multiple frequencies (Hausdorff, Peng, Ladin, Wei & Goldberger 1995), and this property is stable during many different speeds of walking (Hausdorff, Purdon, Peng, Ladin, Wei & Goldberger 1996). To characterize this, the concept of walking (concept meaning the motor commands necessary for walking in this case) could be a limit cycle in state space (and a point attractor in phase space), which is approached by the mental trajectory when one wants to execute that movement. As a trajectory follows the cylinder set around the limit cycle, it has anywhere within the diameter of that tube to be variable. For example, a slight difference in force from muscles used in walking may produce a different stride length, which would be a small deviation somewhere along this trajectory.

Allowing for consciousness to take place continuously through time, and also allowing for it to take place *contiguously* through space, imposes boundary conditions on the possible range of experiences, but also allows the subtle variations of such experience to be almost infinite. One account of embodiment claims that

because we are situated in our environments, we must actively use perception and action in order to create cognition (Barsalou, Breazeal & Smith 2007; Wilson 2002). The number of variables in all of these systems is vast, and thus it may seem unreasonable to assume so many degrees of freedom when consciousness appears to be very similar across different people. To a large degree though, attractors in this sea of cylinder sets could be actual physical objects, people, locations, etc. These external points of cognitive reference are fixed because they are not transformed or altered in any way: there are what they are. The environment is doing a large part of constraining an embodied conscious system. Thus, the place for individual variability and nuance of consciousness rests on the trajectory that one traverses when passing by these external objects.

As previously stated, we have defined awareness as the proximity to concepts in mental state space. Developmentally, for the first several years of life as categories and concepts are defined and learned, the cylinder sets still need to gravitate closely and deliberately to concepts because they are ill defined. To achieve cognition, perception and action, the first attempts to cognize, perceive, and act must involve a deliberate visiting of the concept necessarily because the categories are still have fuzzy boundaries. Awareness will then gradually become more and more acute as a result of the concepts becoming more coherent and thus pulling the mental trajectory closer.

As life progresses, the well-worn consciousness trajectories become more and more refined and more dense. This increase in consciousness allows for one to process the same trajectories, while simultaneously not needing to be quite as aware. The distance from any particular concept or percept can safely increase, resulting in decreased awareness of those concepts and percepts. The benefit is that one can process more in parallel and more rapidly: if one doesn't need their trajectory to veer quite so close, it can take more shortcuts but still pass by all the needed locations to produce adaptive behavior. For example, procedural memory is always accompanied by little or no awareness, but when first learning how to drive a stick shift, the practice is very effortful. The subject of mathematics is at first accompanied by effortful learning and a very poignant awareness, but over time one can simply look at a problem and intuitively know approximately what the answer will be. Various processes will have different developmental timelines, and take different people differing amounts of time to achieve a state of little awareness, heightened consciousness.

An approximation of how these differences and subtleties in awareness and consciousness change over development is shown in Figure 3. According to our framework's definition of consciousness, it can be expected to steadily rise over the course of a lifetime, eventually reaching a state of what is typically referred to as "wisdom," where thoughts can be traversed quickly and intuitively. Awareness,

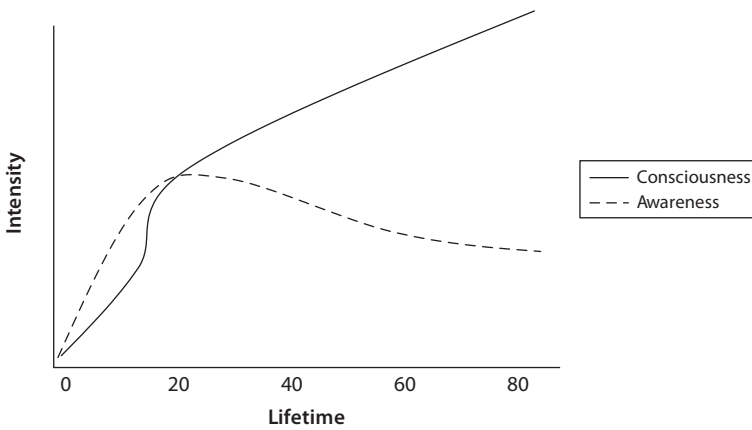


Figure 3. Consciousness and awareness intensity over the course of a lifetime

which we define as an explicit focus on labeled concepts, reaches a peak in early adulthood, and then falls off steadily to various degrees for many years – as the typical trajectories become more streamlined and no longer require the precise visitation of a labeled concept. Computationally, this kind of progression would seemingly be encouraged: if one can process information without taking deliberate, intentional time to process every strand of data, progressing from one thought to the next or one action to the next becomes more efficient. Intuitively, one might expect the progression of awareness and consciousness to be correlated, but the proximity to a concept and the density and smoothness of a cylinder set are actually quite distinct from one another.

A coherence of consciousness, that phenomenal sense of self is in part the trajectories, but is also intensified by the density of information being processed both now, and slightly in the past and the future. Our minds do not simply exist in the present, but rather much of our sense of self is defined by the past, and many cognitive and perceptual operations are able to function as they do because they anticipate the future. The most amount of information is found in the present: what we can see, feel, hear, smell and taste in this moment. But it is also in what we know we are about to see, or what we have seen in the past that defines the present moment. There is a greater density, so to speak, of the phenomenal experience of the now than there is of the past or future. Another way to define what consciousness is may be in essence a distribution of information density over time, such that the present moment contains a peak (Figure 4). One's consciousness is not simply in the present in this case, but defined over all mental time (both memory and anticipation). Non-human animals may have lesser memory and predictive capacities, making their consciousness more hazy, and far less intense and clear

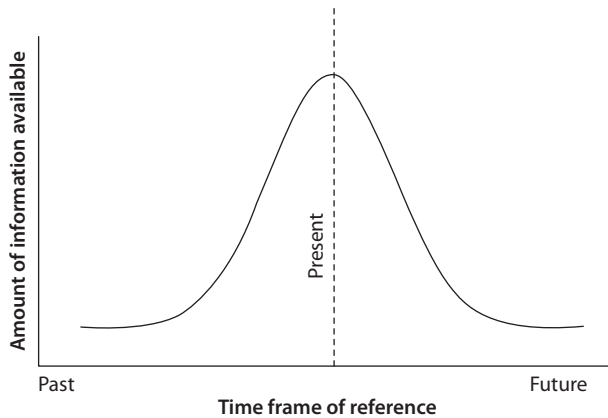


Figure 4. Density of information by time reference

as we experience it as humans. Humans are in this way defined by transitions through time, and lose much of the sense of self when the past is lost (as with retrograde amnesia), or when the ability to track our passage through time and create new memories is lost (as with anterograde amnesia).

3. The pushing and pulling of consciousness

Consciousness has so far been defined as the trajectory along which thoughts and actions flow. Attractors are mathematical descriptions of how concepts pull in or push away the trajectory. Necessarily an accurate characterization of some of the ways in which the trajectory moves in this space must be made.

The way we are using the term “awareness” here treats it as a transitive concept that requires a kind of direct object. That is, to be aware, one must be aware *of something*. This provides for a continuum of intensity. One can be very aware of something, as when the trajectory defining one’s mental state over time visits close to a particular concept, or one can be slightly aware of something, as when the trajectory briefly brushes by the periphery of a concept. By contrast, we wish to allow consciousness to be an intransitive concept that does not involve a direct object. You can be conscious without having to be conscious *of something*.

Consciousness might be best conceived as a degree of statistical orderliness exhibited by the mental trajectory over time. A nearly random skittering trajectory through mental state-space, such as that produced during dreamless sleep perhaps, does not involve consciousness. If that trajectory were carved into a sequence of discrete states and then shuffled like a deck of cards, reconstructing the original

sequence would be nearly impossible, because the states show little or no order in their transitions. However, a smooth and more predictable trajectory that adheres somewhat to the existing cylinder sets of frequently visited sequences of states (e.g. typical thoughts) will be a dynamic pattern that has more predictability and less entropy from time slice to time slice. This statistical orderliness in the trajectory can be seen as an indicator of how conscious that portion of the trajectory is. As certain orderly patterns recur, they form attractor-like ruts in the state-space that become cylinder sets and eventually relatively routinized patterns and sequences of behavior.

As those cylinder sets get formed in the state-space of the mind, some will be frequently visited thoughts that are closely adhering to something like a linguistic sequence (such as “this coffee tastes good”), whereas others will be frequently visited periods of drift in between labeled concepts such as that in Figure 2. The former trajectory carries with it a degree of consciousness (because it is reasonably orderly), and also a level of awareness (in our use of this term) because it has a sequence of labeled concepts stringing the thought together. The latter trajectory (Figure 2) involves consciousness because it is a relatively smooth (non-random) movement through the state space, but less awareness because it is further from the concepts. Although a smooth non-conceptual trajectory like that does not involve a focus on any specific concepts, it nonetheless does impart a sense of being in time, an intransitive consciousness separate from any transitive awareness.

If statistical orderliness and predictability of the trajectory is indicative of consciousness, then it is possible to have a continuum of intensity for consciousness, just as we have for awareness. Importantly, for consciousness to exist there *must* be movement through the state space over time; a non-moving trajectory would mean there is no neural firing. With a highly repetitious looped trajectory (such as walking), or perhaps a stable pattern that hardly moves at all in the state space, this unusually high statistical orderliness (or very low entropy) in the mental state over time might correspond to a heightened consciousness. Meditation is one way in which highly repetitive behavior can lead to something such as a heightened phenomenal experience of awareness or consciousness. In one version of meditation, a nonsense word will be repeated silently again and again as the mantra, inducing a looped mental trajectory. This particular form of greater statistical predictability in the trajectory corresponds to a higher intensity of consciousness, but would decrease awareness because the word is nonsensical.

Other forms of meditation may have differing effects on awareness. Hodgins and Adair (2010) demonstrated that mindfulness meditators, who practice at steering away from labeled thoughts during their meditation, exhibit better performance on change blindness tasks and are less affected by distracting visual cues in attention tasks. Practice at maintaining stability in unlabeled regions of mental

state space allows their attention to not get overly engaged on any one thing and thus they can readily accommodate changes in their environment. That is, mindfulness meditators are more aware of everything because they are less aware of any one thing.

In a brain that functions as a complex dynamical system, with richly interactive neural subsystems (and not informationally-encapsulated modules), there is a multitude of factors that can influence the path taken by the mental trajectory in this very-high-dimensional state space. With bi-directional neural connections between visual, auditory, linguistic, reasoning, memory, emotional and motor systems, just about everything can influence just about anything else (Spivey 2007, Chapter 5). For example, audition influences the mental trajectory for visual perception (Sekuler, Sekuler & Lau 1997; Shams, Kamitani & Shimojo 2000). Vision influences the mental trajectory for speech perception (McGurk & MacDonald 1976) and sentence comprehension (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy 1995). Subtle modifications in visual context can instigate new insight in reasoning (Grant & Spivey 2003). Constraints on motor movement influence visual judgments (Bhalla & Proffitt 1999), and linguistic input influences motor processing (Nazir et al. 2008).

Emotional experience is likely to temper consciousness in very different and non-trivial ways. For example, early experience with abuse distorts how children see faces (Pollak & Kistler 2002). In this study, children who had been physically abused were biased toward categorizing faces as angry, on a sad to angry continuum. Compared to a control population, these children actually perceive the faces differently, seeing angry and sad in a distorted manner compared to non-abused children. Emotional and perceptual experience can alter the state space of consciousness, and intensify an attractor concept (e.g. in conflict: Coleman, Vallacher, Nowak & Bui-Wrzosinska 2007). Emotion is an intense aspect of consciousness and awareness, and can radically alter perception, cognition, and actions. For example, a fact associated with an emotion is more memorable (Cahill, Babinsky, Markowitsh & McGaugh 1995). An emotional memory or fact could be modeled by a deeper attractor basin, making it more easily accessible and stable over time than non-emotional facts. Emotion can moderate not only the intensity of a concept or the speed and degree to which a trajectory moves around such a concept, but it could radically alter the surrounding space (e.g. sacrificing some “sad face” real estate to make room for the more intense “angry face”).

As previously stated, qualia are commonly postulated as discrete bits of information that are difficult to accommodate in the present story, with its emphasis on continuous-time trajectories. An individual quale of what it is like to be in a particular location in mental state-space is a difficult nut to crack. Part of the reason for that is that the trajectory traversed by the mind in its state space is usually quite

smooth and continuous when it is conscious (Spivey 2007, Chapter 12). Therefore, finding a segment of the trajectory that has an unambiguous start point and end point, so that it can be called an individual quale, will be arbitrary wherever one makes the cut. Indeed, one might be safer to treat this “what it is like” notion as a mass noun, such as *qualius*, rather than an enumerable count noun, such as *qualia*. As you read this, you are experiencing some *qualius* right now, as well as some awareness. That said, the presence of *qualius* in a conscious trajectory does not play a causal role in determining its path. It merely provides a residual data stream for later meta-cognitive analysis. Thus, placing *qualia*, or *qualius*, at the center of consciousness research may actually be more of a distraction than a guiding principle (see Edelman 2008, p. 436).

4. Discussion

Psychological and cognitive science has made great progression in the art of isolating individual processes, but an account of consciousness must describe the system at a macro-level of analysis. Give them more than two interacting variables, and most statistical and experimental methods are at a loss. The cylinder set account emphasizes the importance of all relevant variables, and is in itself nothing more than a description of a singular trajectory constantly on the move. At some point, a science of consciousness might need to let go of isolating those one or two variables and determining their causal roles, and simply describe the trajectory.

This chapter has reviewed and described an approach to consciousness where concepts and actions and even the environment are intertwined as sets, containing point attractors (which come and go as the parameters of the system change over time). The attractors often have conceptual labels associated with them, as in the mathematical field of *symbolic dynamics* (see Atmanspacher & beim Graben 2007; Dale & Spivey (2005) and Duch & Dobosz 2011). These labels are what become part of explicit awareness. These attractors function to push and pull a trajectory, along which consciousness flows. As life progresses, the trajectories form cylinder sets that connect the concepts in various sequences. The proximity of the trajectory to a concept is defined as awareness, where over a lifetime awareness is able to decrease on average as the trajectory learns to take short cuts. Due to this streamlining of the cylinder sets, the level of consciousness (i.e. coherence of the cylinder sets and adherence of the trajectory to their central threads) typically goes up over the course of a lifetime.

What it means to be more conscious is in how concentrated experience has become in a singular, integrated whole. The intensity of consciousness relates

to how dense and smooth a cylinder set has become over time. This powerful integrated experience is in the movement of that trajectory. This density and the development of certain aspects of consciousness is likely to be mitigated by both individual differences and differences in culture. Similar experience leads to similar awareness, and the core of that experience, *what it is to be conscious*, is the same.

References

- Atmanspacher, H. & Beim Graben, P. (2007). Contextual emergence of mental states from neurodynamics. *Chaos and Complexity Letters*, 2, 151–168.
- Barbour, J. (2001). *The end of time: The next revolution in physics*. New York, NY: Oxford University Press.
- Barsalou, L.W., Breazeal, C. & Smith, L.B. (2007). Cognition as coordinated non-cognition. *Cognitive Processing*, 8, 79–91.
- Bhalla, M. & Proffitt, D.R. (1999). Visual-Motor recalibration in geographical slant perception. *Journal of Experimental Psychology: Human Perception & Performance*, 25, 1076–1096.
- Cahill, L., Babinsky, R., Markowitsch, H.J. & McGaugh, J.L. (1995). The amygdala and emotional memory. *Nature*, 377, 295–296.
- Choi, J.K. & Kim, S.C. (2007). Environmental effects on gene expression phenotype have regional biases in the human genome. *Genetics*, 175, 1607–1613.
- Coleman, P.T., Vallacher, R.R., Nowak, A. & Bui-Wrzosinska, L. (2007). Intractable conflict as an attractor: A dynamical systems approach to conflict escalation and intractability. *American Behavioral Scientist*, 50, 1454–1475.
- Dale, R. & Spivey, M. (2005). From apples to oranges to symbolic dynamics: A framework for conciliating notions of cognitive representation. *Journal of Experimental & Theoretical Artificial Intelligence*, 17, 317–342.
- Duch, W. & Dobosz, K. (2011). Visualization for understanding of neurodynamical systems. *Cognitive Neurodynamics*, 5, 145–160.
- Edelman, S. (2008). *Computing the mind: How the mind really works*. New York, NY: Oxford University Press.
- Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press/Bradford Books.
- Gandhi, S., Heeger, D. & Boynton, G. (1999). Spatial attention in human primary visual cortex. *Proceedings of the National Academy of Sciences of the USA*, 96, 3314–3319.
- beim Graben, P., Gerth, S. & Vasisht, S. (2008). Towards dynamical-system models of language-related brain potentials. *Cognitive Neurodynamics*, 2, 229–255.
- Grant, E. & Spivey, M. (2003). Eye movements and problems solving: Guiding attention guides thought. *Psychological Science*, 14, 462–466.
- Hausdorff, J.M., Peng, C.K., Ladin, Z., Wei, J.Y. & Goldberger, A.L. (1995). Is walking a random walk? Evidence for long-range correlations in stride interval of human gait. *Journal of Applied Physiology*, 78, 349–358.

- Hausdorff, J.M., Purdon, P.L., Peng, C.K., Ladin, Z., Wei, J.Y. & Goldberger, A.L. (1996). Fractal dynamics of human gait: Stability of long-range correlations in stride interval fluctuations. *Journal of Applied Physiology*, 80, 1448–1457.
- Hodgins, H.S. & Adair, K.C. (2010). Attentional processes and meditation. *Consciousness and Cognition*, 19, 872–878.
- Huttenlocher, P.R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia*, 28, 517–527.
- Ito, M. & Gilbert, C. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, 22, 593–604.
- Jaenisch, R. & Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33, 245–254.
- Kello, C.T., Anderson, G.G., Holden, J.G. & Van Orden, G.C. (2008). The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science*, 32, 1–15.
- Kello, C.T. & Van Orden, G. (2009). Soft-assembly of sensorimotor function. *Nonlinear Dynamics, Psychology, and Life Sciences*, 13, 57–78.
- Lamme, V. & Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience*, 23, 571–579.
- Maye, J., Werker, J.F. & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, 101–111.
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
- Motter, B. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology*, 70, 909–919.
- Nazir, T.A., Boulenger, V., Roy, A., Silber, B., Jeannerod, M. & Paulignan, Y. (2008). Language-induced motor perturbations during the execution of a reaching movement. *Quarterly Journal of Experimental Psychology*, 61, 933–943.
- Nisbett, R. & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Onnis, L., Waterfall, H.R. & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, 109, 423–430.
- Pollak, S.D. & Kistler, D.J. (2002). Early experience is associated with the development of categorical representations for facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 99, 9072–9076.
- Rampon, C., Jiang, C.H., Dong, H., Tang, Y., Lockhart, D.J., Schultz, P.G., Tsien, J.Z. & Hu, Y. (2000). Effects of environmental enrichment on gene expression in the brain. *Proceedings of the National Academy of Sciences*, 97, 12880–12884.
- Saffran, J., Newport, E. & Aslin, R. (1996). Word segmentation: The role of distributed cues. *Journal of Memory and Language*, 35, 606–621.
- Schooler, J., Ohlsson, S. & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology*, 122, 166–183.
- Sekuler, R., Sekuler, A. & Lau, R. (1997). Sound alters visual motion perception. *Nature*, 385, 308.
- Shams, L., Kamitani, Y. & Shimojo, S. (2000). What you see is what you hear. *Nature*, 408, 788.
- Spence, C. & Driver, J. (2004). *Crossmodal space and crossmodal attention* (Eds.). Oxford: Oxford University Press.
- Spivey M.J. (2007). *The continuity of mind*. New York, NY: Oxford University Press.

- Spivey, M. & Spirn, M. (2000). Selective visual attention modulates the direct tilt aftereffect. *Perception and Psychophysics*, 62, 1525–1533.
- Stephen, D.G., Boncoddo, R.A., Magnuson, J.S. & Dixon, J.A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, 37, 1132–1149.
- Stephen, D.G. & Mirman, D. (2010). Interactions dominate the dynamics of visual cognition. *Cognition*, 115, 154–165.
- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, 268, 1632–1634.
- Van Orden, G.C., & Holden, J.G. (2002). Intentional contents and self-control. *Biological Psychology*, 14, 87–109.
- Wells, G. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, 14, 89–103.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9, 625–636.

Two dynamical themes in Husserl

Jeff Yoshimi

University of California, Merced

I describe and partially formalize two aspects of Edmund Husserl's phenomenological philosophy, in a way that highlights their relevance to cognitive science. First, I describe "constitutive phenomenology", the study of structures (what I call phenomenological "models") that constitute a person's sense of reality. These structures develop incrementally over the course of a person's life, and serve a variety of functions, e.g. generating expectations relative to actions, and determining the contents of context awareness. Second, I describe "transcendental-eidetic phenomenology", which posits a hierarchy of laws, each governing the way consciousness must be organized in order for a particular type of thing (a physical thing, a person, a social institution, etc.) to appear.

The German philosopher Edmund Husserl (1859–1938) developed a complex theory of consciousness, which has broad relevance for contemporary cognitive science, and in particular for approaches informed by dynamical systems theory. I reconstruct and partially formalize two broad and forbiddingly-titled areas of Husserlian phenomenology: "constitutive phenomenology" and "transcendental-eidetic phenomenology." Both can be thought of as forms of possibility analysis (analysis of structures in a space of possibilities), and thus both can be understood using concepts from dynamical systems theory, which focuses on possibility spaces or "state spaces". Though I believe these aspects of Husserl can be naturalized in a specific way (Yoshimi 2011), my main purpose here is simply to present Husserl's ideas as clearly as possible, so that they can be drawn on more easily by empirical researchers.¹ In the concluding section I briefly describe how I think these ideas can inform empirical research.

1. There are multiple ways one might "naturalize Husserl", and I don't want to bias this discussion by precluding any particular approach. Two particularly relevant lines of research in the recent literature are associated with (Balduzzi & Tononi 2009) and (Fekete & Edelman 2011). However I make an assumption both groups disagree with (namely, that conscious

“Constitutive phenomenology” is Husserl’s theory of how a person’s overall sense of reality, their “world”, is built up or *constituted* in flowing streams of experience. Think of a kind of internal model developing in layers, as a person has different kinds of experiences in different domains. “Transcendental-eidetic phenomenology” is Husserl’s theory of how, in order to experience particular types of things, certain constraints must be met. These constraints can be thought of as rules governing the way possible experiences must be instantiated if particular types of things are to appear. The theory is a development of themes that go back at least to Kant.

Given space limitations I will only sketch these ideas here. I will, however, try to be clear about what the open questions are and what work remains to be done.

1. Background

A basic construct, from which all other Husserlian constructs can arguably be derived, is a *phenomenological state space*, a space *C* of possible conscious states. Each point in *C* corresponds to a possible human conscious state.² *Conscious states* are understood to be what Gurwitsch (1964) called “fields of consciousness”, which encompass everything a person is aware of at a point in time – thoughts, feelings, the visual field, smells, itches, pains, etc. Various questions arise in relation to the conscious field. For example, it is not clear how extensive it is. Gurwitsch thought we have a fairly rich context awareness, encompassing our sense of our body, some sense of the immediate past and future, etc., while others have claimed that we have a surprisingly limited overall awareness at any time (O’Regan & Noë 2001). There are also questions concerning the notion of an *instantaneous* conscious state. Perhaps there are no conscious states at times, but only over durations. In that case we can redefine “states” (and hence points in *C*) to be brief durations of consciousness.

states at times are determined brain states at times; what is sometimes called “synchronic” mental-physical supervenience).

2. We will focus on conscious states possible for a particular person at a particular “stage of maturity” (more on this below). We can think of this as an accessible subset of *C* for that person, at that stage of maturity. For simplicity, references to accessible subsets are suppressed here, and I simply refer to *C* in general. Also note that these ideas can be generalized beyond human consciousness to all logically possible conscious states. For more on both points see (Yoshimi 2011).

The question naturally arises of what the mathematical structure of C is. Is it a metric space, so that we can say that some pairs of conscious states are more similar to one another than other pairs are? For example, two experiences of a tree while walking around it seem intuitively to be more similar to one another than either of those experiences and an experience of skydiving.³ Is C a vector space, so that we can meaningfully speak of adding conscious states and multiplying them by scalars?^{4,5} These are open questions, though it is typically assumed that C is at least a metric space.

We can represent what William James famously called the “stream of consciousness” as a path in C , a time-ordered succession of conscious states, something like a trajectory for a phenomenological dynamical system. Whether these paths must be continuous is another open question. Can there be phenomenological discontinuities? Perhaps when we go to sleep or wake up there is discontinuity, though perhaps even in those cases we gradually “fade in and out” of a null state. Can a path in C ever cross itself – that is, can one ever have precisely the same conscious experience twice? Husserl denies this, since the temporal context of experience is always changing and so it seems we can’t have the same exact total experience twice.

3. When I say two conscious states C_1 and C_2 are more “similar” to another than another pair of conscious states C_3 and C_4 are, I mean that the distance between C_1 and C_2 is less than the distance between C_3 and C_4 . “Similarity” in this sense may not be the same as judged similarity, because there is evidence that judged similarity cannot be captured by a metric space representation (Tversky 1977). For example, distances are symmetric, but judged similarities are not. One might agree to “The father is like the son” but not “the son is like the father”. This suggests that C is not a metric space. On the other hand, it could be that C is a metric space, but that judgments of similarity are based on a higher order non-metrical process.

4. Some have explicitly argued that C is a vector space or a structure in such a space. For example, (Churchland 2005) uses vector addition in a color perception space to represent chromatic fatigue. (Stanley 1999) describes C as a structure in a vector space, where (modifying Stanley’s ideas a bit), scalar multiplication corresponds to intensification of a state (I see this same tree, but more vividly), and vector addition corresponds to a kind of experiential conjunction (an experience of seeing a tree plus an experience of smelling a rose corresponds to an experience of seeing a tree and smelling a rose). On the other hand there are reasons to doubt whether C is a structure in a vector space, but the issue is complex and I will not elaborate on it here.

5. Below when I speak of C as a vector space, I mean to leave open the possibility that it is a subset of such a space (e.g. a hypercube or some other kind of solid). In that case the vector space operations would have to be suitably restricted.

The space C is large and unwieldy (compare the state space of an entire brain, which tracks, even on a simple connectionist representation, the activity of billions of neurons).⁶ What is typically of more interest in particular phenomenological contexts are specific *types* of experience – color experiences, bodily experiences, cognitive experiences, emotions, etc. Husserl approaches the issue via his theory of parts and wholes or “mereology.”⁷ Husserl uses this theory to consider how conscious states can be broken down into various kinds of parts. One part of my current conscious field is my visual field, and one part of that is my awareness of the laptop before me. I also have some sense of my body, some thoughts, etc. Each is in some sense a part of my total consciousness now. What we’d like is a way to talk about *all possible parts* of various kinds – all possible visual experiences (all possible visual parts of total experiences in C), all possible bodily experiences (all possible kinesthetic parts of total experiences in C), etc. To do so, we can consider *subspaces* of C .

What exactly is meant by a phenomenological “subspace” is yet another open question. One approach is to treat C as an n -dimensional vector space, in which case the concept of a subspace is well-defined.⁸ On this view, C is a product of finitely many subspaces, and we can think of any particular conscious state as having mereological parts which correspond to points in these subspaces. Thus, we can take the space of conscious states C to be a product of a visual experience space V , a bodily configuration space B , a thought space T , an emotion

6. The dimensionality of a system can sometimes be reduced by observing that relevant behaviors of the system lie on a lower-dimensional subspace of its full space. So, the effective dimensionality of “brain space” could be lower than the connectionist representation suggests.

7. Mereology is the study of parts and wholes. Its modern formalization arguably begins with the third of Husserl’s *Logical Investigations* (Husserl 1970). Husserl wanted to distinguish different kinds of part-structures in the phenomenal field. For example, my experience of the door is “part” of my experience of the whole house. My experience of the door’s color is also a part (but in a different sense) of that experience. For a review of contemporary mereology see (Varzi 2011).

8. On the possibility of representing conscious states as vectors in a vector space see Note 4. If C is not a vector space it could still be thought of as a topological space, in which case subspaces are still well-defined. However, because topological spaces cannot in general be written as Cartesian products of subspaces (e.g. in the case of a Möbius strip), some details of the account given here would have to be modified.

space E , and finitely many other subspaces, so that $C = V \times B \times T \times E \dots$ ⁹ The idea is that any particular conscious state has a visual component, a component corresponding to bodily experience, some thought component, some emotional coloring, and so forth.¹⁰ To handle cases where a given conscious state has no component of a given kind (for example, no visual component), we can allow that the zero vector of each subspace corresponds to absence of that type of experience. Presumably some of these subspaces can be further decomposed into lower dimensional subspaces (e.g. the visual experience space might be a product of multiple spaces corresponding to color experiences at each location in the visual field).

A path in C – representing an interval of a stream of consciousness – can be projected to any of C 's subspaces, so that we can consider “sub-streams” of the total stream corresponding to particular aspects of one's unfolding consciousness. For example, if we project a path from C to V , what results is the stream of specifically visual experiences within that stream of consciousness. Each point along a projected path in V corresponds to a visual experience that is a mereological part of the corresponding point in C . When a path is projected from a higher to a lower dimensional space (as in this case), its topological structure can change. For example, a projection can “fold” a path so that it crosses itself, even if it did not previously cross itself (see Figure 1). This fact has phenomenological consequences. Even if Husserl is right that paths in C don't cross themselves (we never have the same total experience twice because of ever-changing context awareness), this does not necessarily apply to paths once they are projected to subspaces of C . For example, we can probably have the same visual experience twice by blinking before an unchanging scene, and we are all but guaranteed to have the same color sensations twice at a given location in the visual field (Yoshimi 2011).

9. I am assuming that an n -tuple $(v_i, b_j, \dots)$ in such a product corresponds to a conscious state which contains v_i, b_j , etc. as mereological parts. Also recall from Note 5 that I leave open the possibility that C is actually a subset of the full vector space, which is useful to note here insofar as some combinations of phenomenological parts may not be possible.

10. This is also consistent with Husserl's own approach to such problems, which emphasizes “spheres” of experience (e.g. the “sphere” of sensory experiences, of affective experiences, etc.). See (Yoshimi 2010).

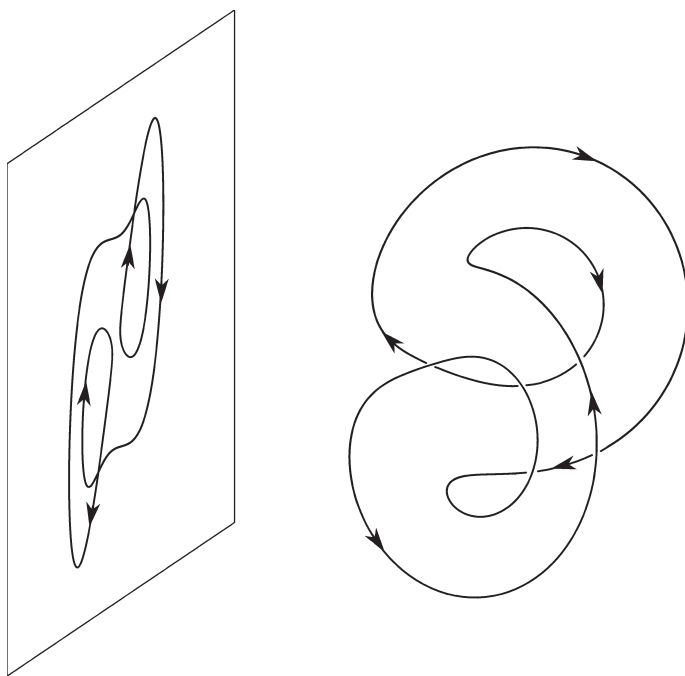


Figure 1. A path in a 3-dimensional space projected to a 2-dimensional subspace. The path in the 3-dimensional space does not cross itself, but its projection to the 2-dimensional subspace does cross itself.

In what follows, it will be important to distinguish C and its subspaces – which are fixed and unchanging relative to a person’s overall stage of maturity¹¹ – from various structures inside of those spaces, which change and morph as a person learns. V contains all possible visual experiences I could possibly have – it is fixed, for the purposes of these analyses. On the other hand, within V there are probability relations between these experiences, whereby, depending on where I am and what

11. By a “stage of maturity” I mean a person’s overall level of development, which determines the range of possible experiences they could have. I am at the same stage of maturity I was last week, but I am at a different stage of maturity than I was when I was 1, 5, or 10 years old (the set of possible experiences for my 5 year old self is different from the set of possible experiences for my current self). Presumably a person’s stage of maturity corresponds to broad anatomical properties of their brain. Thus, if I were to suffer serious head trauma or a massive stroke, what I am calling my “stage of maturity” would change. The purpose of this concept is to allow us to fix the phenomenological state space for a person. Of course, the relevant anatomical properties are themselves gradually changing, so this is ultimately a fiction, but it is a convenient fiction insofar as it facilitates dynamical analyses like these. Similarly assumptions are often made in dynamical systems analyses, though they are rarely made explicit.

I am doing, I expect some visual experiences to occur more than I expect others to. These probability relations change over time, and can be thought of in terms of a changing structure inside of an ambient space which does not change.

2. Horizon theory/constitution

Husserl was, like Kant, a transcendental idealist, who believed that reality could be understood as a kind of projection of consciousness. The exact nature of Husserl's position (or for that matter, Kant's) is controversial, but the interpretive controversies are irrelevant here. Our focus will be on certain formal constructs Husserl used in developing his position, that are useful whether or not one accepts Husserl's broader philosophical views, and that are particularly relevant to research in cognitive science. In particular, we will focus on Husserl's concept of the total "horizon" of our understanding of the world, which in some sense "constitutes" reality as we know it. In this section I develop Husserl's account and show how it can be formalized using the tools introduced above.

A basic feature of our phenomenology is that we assume a real world exists and that we are a part of it. The claim seems too obvious to merit stating. To motivate it, consider the contrast between Husserl's focus on our awareness of a *real world outside of us*, and the prevailing introspectionist psychology of Husserl's day, which emphasized inner awareness of one's own sensory states (e.g. an ability to discriminate tones in a chord or colors in an after-image). Husserl is, one might say, more "outwardly" oriented in his phenomenology than the introspectionists were, insofar as he emphasizes our awareness of the external world. Husserl begins by describing the "natural attitude" of everyday life, a kind of default mode of life which is contrasted with various forms of theoretical reflection. The Husserlian goal is to describe these experiences in the natural attitude as accurately as possible, to capture everyday life *just as it is*. One feature of everyday life, says Husserl, is that we believe the things around us are real – we are naïve realists in the natural attitude. As I sit here writing this, I am at a desk in a room and I assume the desk and room are real things. You reading this no doubt assume that the screen or paper you read these words on is real. Only in the grip of extreme philosophical reflection will someone doubt the existence of the world.

So this is our basic datum: in everyday life we take it for granted that a real physical world exists. Moreover, we have a complex, largely tacit understanding of how this world works. We have a sense of how it is laid out, and where we are in it. We have a sense of how individual things work (a naïve physics). I believe that if I drop my laptop it will break. I believe that I have a physical body that moves in certain ways in this world, and that must be protected from damage. Husserl spent

much of his career developing a theory of this implicit understanding we have of the world around us. Rather than describing it in terms of a list of tacit beliefs, he says that our understanding of the world can be understood in terms of a quasi-mathematical structure (Yoshimi 2009), a “manifold” or “horizon” – what I will call a phenomenological “model” – which corresponds to a kind of picture of a person’s total assumed reality. The job of constitutive phenomenology is to describe this model and to explain how it “constitutes” a person’s overall sense of reality.

Before turning to a formal analysis of horizon theory, here is an intuitive sketch, in terms of a somewhat unusual and mixed set of metaphors. Think of a horizon model as an actual model of the world, spread out on a huge table, like the scale models used in special effects and miniature wargames. Now think of your current visual experience as a translucent, flexible image – one view of the model – that fuses or morphs in to it, becoming part of the model as you have that experience. Your visual experiences over time are successively fused in to this model, and over time the model as a whole becomes more detailed and articulated in familiar areas. In other areas it is sketchy but not empty – even though you’ve never been to North Dakota you still have some sense of what it is like there. If you were actually to visit North Dakota, you would begin to flesh out that part of the model in more detail. Historical maps of the world – detailed in Europe but sketchy in the American interior – come to mind. But this is no static map, it is a dynamical model that becomes more articulated as a person visits different parts of the world.

I now turn to a more formal analysis of Husserl’s horizon theory, whereby these models corresponds to structures in subspaces of *C*. I will not attempt an exhaustive analysis of Husserl’s theory here, but rather will focus on three main features. (1) A person’s horizon-model corresponds to a totality of possible experiences, including every experience a person has had at that stage of maturity. (2) This horizon is more or less “delineated” or “articulated” in different regions. (3) As a person has new experiences the horizon is updated. I will focus on visual and bodily experience – which, says Husserl, constitute the physical world as we know it (Husserl 1997) – and will then briefly consider extensions to other domains of a person’s sense of reality.

First, a manifold or horizon is a *totality of possible experiences* of a given kind for a person, what Husserl also refers to as “pre-knowledge” that informs our current experiences.¹² Husserl focuses on possible experiences of individual

12. The term “horizon” has several distinct meanings in Husserl, all of which I draw on here. (1) a substrate of background knowledge similar to what (Searle 1983) calls the “background of intentionality”, which captures all of a person’s tacit knowledge of the world, (2) a structure in a possibility space, what I have called a phenomenological “model”, (3) a “felt horizon” or some more specific “feeling” of context. When I say “horizon” without qualification I mean sense (2).

objects (what he sometimes calls an “internal horizon”), a “manifold of percepts belonging to the same external thing” (Husserl 1970, p. 307), but he also generalizes this in various ways, e.g. to all possible experiences of a physical world. I will focus on our model of the physical world as it is visually apprehended, which corresponds to a subset of V , the space of possible visual experiences.¹³ This subset contains, at a minimum, all actual visual experiences of the world you have had within your present stage of maturity. Imagine taking the stream of experiences you have had during this period, a winding path in C , and projecting it to V . This projected path doubles back on itself many times in some regions of V , but has barely ventured in to other regions. The resulting set of points in V does not, however, correspond to the complete set of possible visual experiences of a person’s world. In particular, it contains experiences a person has not actually had, but that are nonetheless consistent with that person’s overall understanding of the physical world. The total set of visual experiences a person could have of the physical world as they understand it, is their horizon-model of the world.¹⁴

Second, the horizon is more or less “articulated” or “delineated” in different regions. Husserl also describes this in terms of the expectations (what he calls “motivated” or “prefigured” possibilities) we have in different circumstances. When I am in my home town, I have very specific expectations about what I will see next when I make turns on various streets; but when I drive through an unfamiliar town I have much less specific expectations. These ideas can be formalized in terms of a function which associates my current visual experience and bodily movement with what visual experiences I expect to have next. Husserl himself describes such a function:

When I undertake a series of movements in the free system, “I move myself,” the appearances that are arriving are already prefigured. The appearances form dependent systems.... Only through this interplay of independent and dependent variables is what appears constituted as a.. perceptual object....

(Husserl 2001, pp. 51–52)

13. It does not correspond to all of V , because some visual experiences correspond to random sensory data – like “snow” on a tv – and thus do not present a world. Cf. the discussion of maelstroms in Section 3.

14. The question arises of how to generate the full horizon for a person from those points in V that a person has actually instantiated. One approach is to try to find a mathematical manifold (or finite set of manifolds) that these points lie on or near. Such a manifold would embed all actual points and contain other points “between” these as well. Finding such a manifold is non-trivial, but corresponds to an active area of research, sometimes referred to as “manifold learning” (Izenman 2008).

Notice that Husserl refers to the “appearances that are arriving” as dependent variables, which are a function of independent variables which include bodily movement. We can formalize this in terms of an “expectation function”, which takes three arguments: (1) a person’s overall background knowledge (see Note 12), (2) their current visual experience in *V*, and (3) a felt bodily movement in *B*. Its output is a probability density function on *V*. The probabilities obtained from this density function can then be interpreted as degrees of surprise, where lower probabilities correspond to greater surprise. If I am before a wall and believe there is a window behind me, and I turn around, some resulting visual experiences will surprise me; others will not. If I turn around and see the window I expected to see, I won’t be surprised. On the other hand, if I see no window at all, or a large gaping hole in the wall, I will be very surprised. These correspond to more or less probable visual experiences relative to the probability density function produced by the expectation function. This in turn allows us to make sense of the notion of more or less “articulated” regions of the horizon-model. In “well articulated” regions of the horizon, the expectation function will produce relatively sharp density functions insofar as we have specific expectations relative to different movements; in less well-delineated regions, the function will produce flatter density functions insofar as we have less specific expectations relative to different movements.

Finally, the horizon is constantly being updated; as Husserl says, it is always in “flux.” Husserl describes this flux in several ways. On the one hand, he refers metaphorically to the acquisition of “sediments” or “precipitates”: as we interact with the world our understanding of it is enriched, in a kind of geological process whereby sediments of knowledge are deposited layer-by-layer on to our current understanding. Less metaphorically, he describes this flux in terms of changes to what I am calling the expectation function defined on a horizon. In particular, he describes a kind of associative learning rule, whereby expectations that are confirmed or “fulfilled” on the basis of our ongoing experiences are strengthened, and those that are “frustrated” are weakened (he also refers to incremental changes in the “weightings” of associative links). This can be formalized by viewing updates following expectation fulfillment or frustration as changes in background knowledge, which in turn alter the probability density functions produced in response to a given action-perception pair. Every time I see something I expect to see, background knowledge is updated in such a way that the probability distribution incrementally sharpens around that expectation. If I turn a familiar corner and see the restaurant that is always there, my confidence in seeing that restaurant is incrementally strengthened, so that next time I turn that corner I will be just a little more confident about seeing it again. Conversely, when my expectations are frustrated, the probability distribution is changed so that whatever I did in fact see becomes more probable. If I turn the

corner and am surprised to see the restaurant has been demolished, I will update my background knowledge, so that next time I turn that corner I will now expect to see the demolished restaurant.

Having outlined and partially formalized Husserl's horizon theory, let us briefly consider some of the work it does in his broader phenomenological theory.

A first function of horizon models is to generate feelings of familiarity. In areas of the horizon-model where the expectation function generates relatively sharp density functions relative to body movements, our surroundings feel familiar. By contrast, in areas where the expectation function generates flatter density functions, our surroundings feel less familiar.¹⁵

A second function of horizon models is associated with (Gurwitsch 1964), who describes fields of consciousness (points in C) as complex structures involving a focus of attention (or *theme*), as well as a *thematic field*, a form of context awareness involving ideas related to the theme.¹⁶ Gurwitsch says that focal experiences are always surrounded by a thematic field.¹⁷ Items in the thematic field can subsequently be focused on, and are thus *potential themes*. For example, if I am looking straight ahead and see a tree, I have some context awareness of the things around the tree, which I could subsequently focus on. The contents of this context awareness seem to be determined by a person's horizon model of the world, and in particular by the most probable visual experiences generated by the expectation function relative to movements in various directions. What I am focally aware of visually corresponds to a current point in V, and my context awareness of what surrounds what I am focally aware of corresponds to what I expect I would see were I to move my body in various directions.

I have thus far focused on a person's internal phenomenological model of *physical* reality, insofar as it is constituted by visual experiences and bodily movements. But one's sense of what exists is more complicated than that. Our overall sense of reality includes many kinds of things : our understanding of friends and

15. Relevant empirical literature includes the psychological literature on familiarity (Yonelinas 2002), and on skill acquisition and expert performance (Ericsson & Lehmann 1996). Some related philosophical discussions include Searle's account of the background of intentionality (Searle 1983) and Dreyfus and Dreyfus' model of skill acquisition (Dreyfus 1980).

16. He also describes a third structure, a "margin" of subjective data that are not at the focus of attention, and that are not related to the theme.

17. A similar, empirically grounded account, with what I think is a plausible story about neural correlates, is in (Koch 2004).

social groups, of mathematics, of history, of various political controversies and current issues, of fictional characters living in fictional worlds, etc. In each case some version of the whole story told above obtains. In each case it seems we build up a kind of model of the relevant domain over time, and that this model contextualizes our experiences and expectations when we have experiences in that domain. Consider some examples: thinking through a mathematical problem, debating the nature of justice, planning a trip, gossiping, discussing a policy, or evaluating a film. In each case one's moment-to-moment thoughts are embedded in an explorable context of other thoughts and experiences one has had and could have in relation to the topic. In each case there is a distinction between more or less articulated regions of the model (the math I know well vs. the math I hardly know at all; the people I know well vs. people I hardly know), and in the more articulated regions one will feel more confident and familiar than in the less articulated regions. Moreover, just as context awareness when exploring the physical world is determined by a horizon model, so too, it seems, with context awareness when exploring non-physical domains. To take one of Gurwitsch's favorite examples, when engrossed by a mathematical problem "we may experience references to the possible solutions of the problem... directions in which a solution might be found... [or] assumptions and theorems which might seem to assist in solving the problem" (Gurwitsch 1964, p. 1). However, there are tricky issues here. When I am solving a math problem, debating a political issue, or thinking about my friends, what are the possible "actions" relative to which expectations are generated? How does emotional valence figure in to this story? Places where bad things happen seem to acquire a negative tinge, and similarly for places where good things happen, but it's not clear that emotions have their own horizon structure independently of a kind of overlay on other constituted domains. These are, I think, deep and interesting topics, which I hope to explore in future studies.

3. Transcendental-eidetic phenomenology

A second form of possibility analysis in Husserl is associated with what he variously calls "eidetic," "transcendental" or "transcendental-eidetic" phenomenology. The word "eidetic" is a modification of "*eidos*" or "essence", an unchanging and universal truth (the reference is ultimately to Plato, whose theory of Forms emphasized a realm of pure forms). "Transcendental" is a reference to (Kant 1999), who famously emphasized conditions on the possibility of various kinds of experience. For example, some conditions on the possibility of judging *that a brown table is before me* are, according to Kant, that I be able to perceive things in space and time,

and that I understand what extensive magnitudes, intensive magnitudes, and relations are. Husserl combined these ideas in his transcendental phenomenology. He aimed to describe a set of pure, unchanging “laws” of consciousness, in the form of a set of essences, each of which corresponds to a condition on the possibility of a particular type of experience.

Husserl’s method is to identify essences or laws governing particular types of conscious experiences, using a method called “free variation.” To understand what these essences are and in what sense they “govern” types of experience, it helps to begin by asking what would happen if there were no essences or laws of consciousness at all. If we allowed that a person’s experiences could have any arbitrary character, then subjective life could simply be a procession of chaotic sensory data, what Husserl calls a “maelstrom.” In such a case we would have no experience of a *world*, no sense of an enduring reality, and perhaps no sense of being persons at all. As Husserl says, “it would be a maelstrom so meaningless that there would be no I and no thou, as well as no physical world – in short, no reality” (Husserl 1997, pp. 249–250). If we are to experience a world – if we are to have a sense of an enduring reality – then subjective data must be *organized* they must hang together and unfold in certain coherent ways. Elaborating these rules – rules that must be followed if a world is to appear – is the project of transcendental-eidetic phenomenology.

Another way to think of these rules in terms of the movie, *The Matrix*. Recall that the people in the *Matrix* are in casings, and that electrical impulses are pumped in to these casings to produce the illusion of a world. The important point for our purposes is that someone had to program a computer to deliver these electrical impulses in an appropriate way, producing meaningful inputs and responses to motor outputs (for example, when an encased person sends out signals for moving right, an appropriate image should be sent back to her retina). That is, someone had to program a computer so that it would make the person in the casing believe she was in a world. Someone had to *program a world*. This person, we discover in the third installment of the *Matrix* series, is called “the architect.” The architect had to program rules to prevent people in the *Matrix* from experiencing chaos, he had to force sensory data to appear and to change in response to what people do in a coherent way. These rules are like the rules that Husserl was interested in discovering via his eidetic method.

In practice, Husserl did not focus on high-level, generic rules that prevent subjective chaos altogether (he considered these just to make the general point that having a sense of reality implies the existence of some rules). Rather, Husserl focused on particular types of consciousness, what I will call “domains” D_i , and then asks, what is a condition on the possibility of having an experience or

experiential process in D_i ?¹⁸ That is, what structures are common to all members of a given domain D_i ? Moreover, he believed these domains formed in to a hierarchy, each governed by its own eidetic laws. At the highest level are rules which prevent chaos altogether. At lower levels (some of which correspond to special domains Husserl called “regions”), there are rules governing particular types of experience, what Husserl calls “regional axioms” or “syntaxes” (Yoshimi 2009). Thus we have conditions on the possibility of experiencing visual scenes, unchanging environments, rotations in unchanging environments, interactions with other humans, etc. Continuing the *Matrix* analogy, these laws are like modules of the *Architect*’s program, subsets of the code that govern particular types of appearances: the “physical reality” module, the “other person” module, the “form from motion” module, etc.

Husserl identifies these regional essences by the method of “free” or “eidetic variation.”¹⁹ Free variation can be broken down in to three steps. First, you specify some domain D_i of interest, for example, possible perceptions, possible color experiences, possible conversations, and so forth. Second, you consider an arbitrary exemplar of that domain, and then imagine permutations of that exemplar (being sure to remain within the bounds of the domain in question). In the third and final step you observe what properties $x_1 \dots x_n$ remain constant or “invariant” while you permute the exemplar. These invariants are the “essences” or “eide” of D_i , and they can be thought of as phenomenological laws: for any experience in D_i , it will have the properties $x_1 \dots x_n$.²⁰

There are clearly problems with this approach. In particular, it is not clear how a necessary truth governing a domain of possibilities can be established when one can only consider a proper subset of those possibilities in a finite amount of time. One response to this and similar worries is to simply reject the stronger (and in my view, less plausible) aspects of Husserl’s epistemology and treat his results as fallible hypotheses. In fact, there is evidence that Husserl himself thinks of his

18. By a domain I I mean some structure constructible from C . In some cases D_i is a subspace of C ; in other cases it is a subset of a subspace of C ; in yet other cases it is something more complex, e.g. a set of possible paths in a subspace of C .

19. For a critical review of the literature on free variation see (Kasmier 2003). Also see (Yoshimi 2007).

20. In framing this method, Husserl was influenced by his previous work in mathematics. In fact, his dissertation was on the calculus of variations, a mathematical technique which has affinities with the method of free variation; see (Yoshimi 2007).

project as being as fallible as any human endeavor.²¹ From this standpoint the eidetic method provides a way of framing conceptual/phenomenological hypotheses that are subject to confirmation, revision or rejection based on results in other areas.²² It should also be noted that phenomenology, understood in this way, is not much different from empirical psychology.²³ In both phenomenology and in empirical psychology (and the cognitive sciences more broadly), subjective data and other forms of data can jointly support fallible hypotheses.²⁴

Let us consider some examples to see better how Husserl's eidetic method worked in practice.

Let D_p be the set of perceived physical things. What are the essences of perceived things? What is invariant in this domain of C? What is required in order to experience a physical thing? Take an arbitrary member of this set, say, a ball of wax, an armchair, or a sofa, and imagine variations on it. Let's begin with an armchair. Now change it in some way, for example, change its color – it is a red armchair, a blue armchair, etc. Now change its size, or even let it morph into something altogether different (Husserl is said to have once freely varied a seminar participant into a ball of string). You can permute the imagined object as much as you want, so long as you remain in the set of perceived physical things (so you cannot permute it into something abstract, like a number). No matter how much you

21. As he put it in a lecture course: "phenomenology's scientific articulation, like any theorizing... comes with its own wellspring of deceptions" (E. Husserl 2006, Section 22). For further discussion of Husserl and fallibilism see (Hopp 2009).

22. Compare (Mangan 1991) on "convergent phenomenology": "The virtue of emphasizing a convergent approach to the study of consciousness is that even if each line of evidence is in itself just plausible, the total ensemble, when it converges on the same finding, can give that finding strong support. The principle is the same in certain kinds of radiation therapy: many relatively weak beams of radiation, converging from different angles, produce an intense dose of radiation" (83–83).

23. Husserl himself did not think of phenomenology in this way. He clearly thought that phenomenological data has priority over (and is in fact foundational for) the empirical sciences (Yoshimi 2010).

24. Shimon Edelman (personal correspondence) gives the example of perception science: "All of perception science is rooted in phenomenology in the sense that subjects must rely on their phenomenal states when responding in an experiment... [for example], a subject may be faced with a split display, each half showing a color, the task being to adjust the color of one of the halves so as to match that of the other half." So, in a sense phenomenology is just a kind of psychology. However, there are differences of emphasis and, moreover, Husserl's theory of consciousness and his methodological procedures are, *de facto*, novel. So there is plenty to mine from Husserl's philosophical output. From this perspective the goal of Husserlian neurophenomenology is to integrate his best insights in to the cognitive sciences.

change this exemplar, it will remain extended in space, and subject to change.²⁵ Moreover, no matter how much you deform and change it, you will never be able to see more than one side of it at once. Let us call this the “one-sidedness” of perception. Finally, no matter how crazy a perceived physical thing you create in your imaginings, you will never come up with something which has an inside larger than its outside. Let us call this “containedness.” Thus, extension, changeability one-sidedness, and containedness all seem to be essences of the class of experienced physical things. We can state these as (again, fallible) laws: if you have an experience of a physical thing, that thing will be experienced as extended and changeable as being perceivable from one side at a time only, and as having an outside larger than its inside.

In some cases we can frame essential laws that relate one class of experiences to another. For example, we can consider how the set of perceptual representations of physical things (D_p) is related to the set of imaginative representations of physical things (D_i).²⁶ Husserl suggests that the two classes are related by a 1-1 map. First, he argues that each to each perception in the set of perceptions D_p there corresponds an imagination in the set of imaginations D_i . As he puts it: “there exists an ideal linkage between perception and imagination which assures us *a priori* that to each percept a possible image corresponds” (Husserl 1970, p. 327). He also claims that imaginations can be uniquely associated with perceptions: “every phantasy [imagination] has the value of a possibility, and it guarantees a perceptual possibility” (Husserl 1997, p. 252). To show this, two variations can be performed. We first consider an exemplar in D_p (possible perceptions), in which a thing is directly given, “in the flesh.” For example, my perception right now of this computer. It is certainly the case that I can also close my eyes and imagine the same computer. Now I permute that perception so that the computer is larger or smaller, or viewed from different angles, and in every case I can imagine the same thing. So it is an essence of perception that to every perception some imagination corresponds. Now consider an exemplar of D_i , the set of imaginations. Take any imagination you could have. Say, you imagine a dragon breathing fire. It would not be easy to produce, but in principle we could set things up so you actually perceived the dragon breathing fire. And as we change the dragon in imagination it is always logically possible that the imagined dragon could be perceived in the

25. Husserl borrowed these particular examples of essences from Descartes, who made them famous via his wax example.

26. Both domains contain representational contents abstracted from variations in attention, intensity, emotional coloring, and so forth.

flesh. Similarly with any imagined physical thing. These two essences suggest a 1-1 map between imagination and perception.

A third example shows how Husserl's ideas are related to ideas independently developed in empirical psychology. Let D_B be the class of perceived borders, and consider what is essential to such borders. First consider an arbitrary perceived border in D_B , and then permute that border (being sure that it remains a border through the permutation). Throughout the variation we note that there is always a sudden increase in color change on either side of the border (otherwise we will perceive gradations of color but no sharp boundaries):

A border, indeed, a line results from the leap in quality that occurs... when both parts having been evenly colored... are set off in relief from one another through sharply contrasting colors. A "gradual transition" through a gradually changing qualitative gradation yields a border only if the transition first goes very slowly, and then proceeds very quickly, and then very slowly again.

(Husserl 2001, Section 31)

He also describes this in terms of the rate of color change as a function of space, which suggest a law to the effect that if a boundary is perceived, then the rate of color change at the boundary must be greater than the rate of color change on either side of the boundary. Thus it is part of the essence of perceived boundaries that the rate of color change on either side of the boundary is less than it is at the boundary. This is what allows us to perceive individual objects at all, and by extension, groups of objects, which may be why Aristotle said: "number is perceived by the negation of continuity" (Aristotle 1961, bk. 3). Husserl is here stating, on phenomenological grounds and in an incipient way, ideas that have since been experimentally investigated and confirmed (Frome, Buck & Boynton 1981).

A fourth example involves a more complex domain. In his study of the constitution of space in *Thing and Space* (Husserl 1997), Husserl considers various forms of bodily movement, which constitute different kinds of spatial structures. For example, he considers the "cyclical manifold" of turning, and says that, if we assume an unchanging environment, then if we rotate in place, the sequence of visual images will repeat (that is, we will have a topological circle in V). So, our domain D_T consists of paths in C , which correspond to turning around once in a stationary environment. Now take an arbitrary exemplar of this domain, e.g. spinning in place in your current environment, while assuming that this environment does not change. If we imagine various changes to the rotation, e.g. in terms of its speed, or what objects are around us, we still get a repeating sequence of visual images. We might call this the "periodicity of visual images." This periodicity is an essence of D_T . A projection of any path in D_T to V is, on this account, a topological circle.

The eidetic laws governing expectation generation (discussed in Section 2) – what Husserl also calls “motivational laws” – provide a final example. We already saw how these laws work above, but let us see how they correspond to essences discoverable via free variation. As an illustrative example, we can focus on a domain D_S , of perceptions of a stationary physical environment after having looked left and right.²⁷ More specifically, we can take D_S to contain paths in C that involve a look to the left followed by a look to the right in what one takes to be a stationary environment. Take an arbitrary exemplar of this class: just look to your left and right and then look forward again. Suppose I see brown walls on both sides. At this point we have the following two “motivated possibilities”: if I look right again, I expect to see a brown wall; if I look left again, I expect to see a brown wall. The brown wall would not surprise me, relative to my current visual experience and background knowledge. Thus, there is an eidetic law that says something like this: any time you look left then right in a stationary environment, you will expect to see what you just saw to the right if you look right again, and similarly if you look to the left again. More formally: any continuation of a path in D_S that involves a subsequent look left, will also involve an expectation of what was previously seen to the left, and similarly for a subsequent look right. It is impossible to look left and right in what one takes to be a stationary environment, and *not* expect to see the same thing again if one looks left or right again. The point seems to hold if we consider different environments, different head rotation velocities, etc. So, it seems to be an essence of D_S .

4. Conclusion

I will conclude by briefly describing my approach to naturalizing constitutive and eidetic phenomenology, which is based on the assumption that brain states can be associated with conscious states via a “supervenience map”. This map can be used to study the structure of C , by projecting back and forth between it and the space of brain states B (Yoshimi 2011). In this kind of a framework, horizon structures can be studied using neural network models of embodied brains. The state space of such a model is an approximation of the state space of a real brain. In provisional studies I have embedded these models in virtual environments, and recorded all their states as they interact with an environment. In this way I can

27. The laws described here are instances of a more general set of laws governing arbitrary movements in various kinds of environments. Husserl himself considers movements of the eyes, head, and body, first in stationary, then in changing environments (Husserl 1997).

generate simulated “life paths” in an agent’s state spaces. I then try to find a manifold on which these paths are embedded, which can be taken to be the agent’s “horizon” model of its environment. I can also, at least in principle, compute the expectation function described above, and watch in real time as expectations are generated while the agent explores its environment. In these ways it is possible to visualize and mathematically analyze the kinds of horizon-structures Husserl only describes in a qualitative way.

Transcendental-eidetic phenomenology is also relevant to empirical research in cognitive science, as some (focusing on Kant) have recognized (Churchland 2002; Palmer 2008). In particular, if there are constraints on possible experience in a given domain, those constraints should (assuming supervenience) have some analog in the corresponding domain of possible neural activity. That is, structures constraining activity in subspaces of C should be correlated with structures constraining activity in corresponding subspaces of B.²⁸ This in turn suggests that Husserl’s and Kant’s conditions on the possibility of coherent experience should be associated with constraints on neural processing, which themselves derive from the structure of the brain and its environment. That is, eidetic laws should ultimately be associated with structures in the brain and its environment. Conversely, it could be possible to identify new phenomenological constraints – that is, new eidetic laws – by studying constraints imposed on neural processing by the brain and its environment.²⁹

References

- Aristotle. (1961). *Aristotle, De Anima. Edited, with introduction and commentary.* (W.D. Ross, Trans). Oxford: Clarendon Press.
- Balduzzi, D. & Tononi, G. (2009). Qualia: The geometry of integrated information. *Plos Computational Biology*, 5(8), e1000462.
- Churchland, P. (2002). Outer Space and Inner Space: The New Epistemology. *Proceedings and Addresses of the American Philosophical Association*, 76(2), 25–48.
- Churchland, P. (2005). Chimerical colors: some phenomenological predictions from cognitive neuroscience. *Philosophical Psychology*, 18(5), 527–560.
- Dreyfus, S.E. (1980). *A five-stage model of the mental activities involved in directed skill acquisition.* DTIC Document.

28. See the discussion of “Subspaces and applications” in (Yoshimi 2011).

29. I am grateful to the editors, Tandra Ghose, Scott Hotton, David Vinson, and an audience of graduate students and faculty at UC Merced, for helpful comments and discussion relating to this paper.

- Ericsson, K.A. & Lehmann, A.C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, 47(1), 273–305.
- Fekete, T. & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*.
- Frome, F.S., Buck, S.L. & Boynton, R.M. (1981). Visibility of borders: separate and combined effects of color differences, luminance contrast, and luminance level. *JOSA*, 71(2), 145–150.
- Gurwitsch, A. (1964). *The Field of Consciousness*. Pittsburgh, Duquesne UP.
- Husserl, E. (1970). *Logical Investigations*. (J.N. Findlay, Trans.) (Vol. 1). London and Henley: Routledge.
- Husserl, E. (1997). *Thing and Space: Lectures of 1907*. (R. Rojcewicz, Trans.) (Vol. 7). Springer.
- Husserl, E. (2001). *Analyses Concerning Passive and Active Synthesis: Lectures on Transcendental Logic* (Vol. 9). Kluwer Academic Print on Demand.
- Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Verlag.
- Kant, I. (1999). *Critique of Pure Reason*. (P. Guyer, Trans.). Cambridge Univ Pr.
- Kasmier, D. (2003). *Husserl's Theory of A Priori Knowledge: A Response to the Failure of Contemporary Rationalism*. University of Southern California.
- Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach*. Roberts & Company Publishers.
- O'Regan, J.K. & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5), 939–972.
- Palmer, L. (2008). Kant and the brain: A new empirical hypothesis. *Review of General Psychology*, 12(2), 105–117.
- Searle, J.R. (1983). *Intentionality, an Essay in the Philosophy of Mind*. Cambridge Univ Pr.
- Stanley, R.P. (1999). Qualia space. *Journal of Consciousness Studies*, 6(1), 49–60.
- Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4), 327.
- Varzi, A. (2011). Mereology. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2011.). Retrieved from <http://plato.stanford.edu/archives/spr2011/entries/mereology/>.
- Yonelinas, A.P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, 46(3), 441–517.
- Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, 6(3), 271–291.
- Yoshimi, J. (2009). Husserl's theory of belief and the heideggerean critique. *Husserl studies*, 25(2), 121–140.
- Yoshimi, J. (2010). Husserl on psycho-physical laws. *New Yearbook for Phenomenology and Phenomenological Philosophy*, 10, 25–42.
- Yoshimi, J. (2011). Phenomenology and Connectionism. *Frontiers in Theoretical and Philosophical Psychology*, 2, 1–12. doi:10.3389/fpsyg.2011.00288.

Desiderata for a mereotopological theory of consciousness

First steps towards a formal model for the unity of consciousness*

Wanja Wiese & Thomas Metzinger
Johannes Gutenberg-Universität Mainz, Germany

1. Introduction: Unity, holism, and temporal continuity

In this chapter we present analytical building blocks for the construction of a more formalized theory of conscious experience (Section 2). We then proceed to a first, critical application by arguing that it makes no sense to speak of the “duration of an experience” as a clearly determinable property (Section 3). Finally, we will propose a positive model for a representational architecture that satisfies some of the constraints constituting our general epistemic goal, namely, a richer and more developed theoretical framework for understanding the unity of consciousness (the GSE-architecture; see Section 4). Our general approach is what has been called the “method of interdisciplinary constraint satisfaction” (MICS) by Josh Weisberg: Simultaneously considering constraints at a number of different levels of description, the goal being to use information and techniques from a variety of sources to triangulate on a working concept of consciousness (Weisberg 2006; Metzinger ²2004). Here, what we are interested in is how such constraints can be *integrated* across different levels, yielding a single formal framework that can then integrate diverse theoretical approaches or even existing data-sets. Phenomenologically, we want to focus on the complex relationship between unity, holism, and temporal continuity. Therefore, in this first, introductory section, we will confine the domain of inquiry by setting five phenomenological constraints:

- (GU) Phenomenological constraint #1: When asked, most subjects describe their phenomenology as *globally unified*. We take this to be a special case of

* We wish to thank Shimon Edelman and Kenneth Williford for a number of very helpful commentaries and suggestions.

coherence (on the functional level of description) and, respectively, of *wholeness* (on the phenomenological level of description). It is special by being of a global kind: The overall phenomenology is described as globally unified, because it has unified parts (and not because it is itself part of a unified whole).

- GU is a good candidate for a necessary condition for the ascription of consciousness, because there seem to be no phenomenological configurations, in which it is not satisfied. To be sure, there exist various well-documented breakdowns of coherence on subglobal levels of granularity (as in anosognosia, apperceptive agnosia and a host of other neuropsychological disorders), and it also makes sense to describe *degrees* of global coherence (see also Mroczko-Wąsowicz forthcoming). However, global unity seems to be the most general phenomenal property: Whenever phenomenal experience emerges, this property is always instantiated. Because of its unique character as the *single* most general and *highest-order* form of integration, it is best characterized as a distinct form of “oneness” or unity. All other forms of coherence can be multiply instantiated, in parallel, but only here do we find the singularity which constitutes the subjective experience of living in one, single world. In this world multiple objects, agents, situations and contexts may interact, but the singularity of the highest-order whole becomes a distinct feature of the accompanying phenomenology.

In addition, it is important to note that phenomenal coherence has no direct connection to the coherence of intentional contents (i.e. it is not a *semantic* kind of coherence linked to rational thought and unified agency as, for example, in Schechter 2010:4; although it can certainly go along with it). The kind of global unity we are aiming at is a purely phenomenal property characterizing the fundamental structure of experience, and it could exist in dreams or other delusional states, i.e. in the absence of bodily or mental agency and without any rational form of cognitive processing. Its functional-level counterpart is a dynamic form of *causal* coherence.

- The GU-constraint approximates a highly general feature, because *all* states of consciousness described by human subjects are described as globally unified; consequently GU is blind to a whole series of theoretically relevant features which we subsume under the following constraint.
- (H) Phenomenological constraint #2: Most conscious subjects of experience describe the overall situation or world they experience as characterized by an all-pervading quality of holism, as a phenomenal Holon (for a more extensive description, see Metzinger ²2004, Section 3.2.4). GU is a special case of H by being the *highest-order form* of holism by which a subject’s phenomenology can be described; in addition, it can only be ascribed to the totality of conscious

experience, i.e. to the *whole* phenomenology (and not to parts of it). This means that, in a single subject, this property is never multiply instantiated.

- There is a distinct *phenomenology of holism*, which supersedes the phenomenology of unity: Global unity is actually experienced as the result of dynamic, mereological relations between subglobal wholes; my subjectively experienced world is continuously constituted by flexible hierarchies of part-whole relations. This dynamic internal structure endows the overall reality-model with a quasi-organic structure; it turns it into a “lived” reality, because we experience not only global integration *per se*, but degrees of interdependence, reflecting some aspects of the functional process by which this global holism is constituted through continuous, internal interconnectedness (see constraint 4; STC).
- *Global holism*: As pointed out when introducing GU, global unity (“oneness”) is, on the phenomenological level of description, a special case of “wholeness”. The most general phenomenal property can now be more informatively described as a global form of wholeness, because the different types of phenomenal content constituting our overall subjective experience at any given moment are not integrated like decontextualized atoms or individual members of a set are integrated, but more like parts of a dynamic, overarching whole evolving over time. Internally, this richer kind of integration (to be described below) is of a distinctively non-compositional and non-conceptual¹ nature and its strength fluctuates over time.
- *Convolved holism*: “Oneness” would be a digital, “on-off” property; but phenomenal “wholeness” comes in degrees, it is clearly a gradual property. In order to understand the gradualness of global coherence, one needs a conceptual model of how *subglobal* entities can contribute to the overall effect. The extent to which a given conscious experience is globally coherent corresponds to the extent to which it has *whole* parts, i.e. parts that are connected to other parts in a certain sense. The interesting additional feature (which we here call “convolution”) is that subglobal whole parts can be embedded into each other. Hence, it also makes sense to talk about *subglobal holism* which corresponds to the extent to which a given part itself has whole parts. And again, the only entity which cannot be embedded into a whole part is the global

1. “Non-conceptual” means that it does not result from a mental form of predication or any other inner sort of concept use/concept formation. There is no inner form of agency or top-down causality involved, just a bottom-up process of dynamical self-organization.

conscious model of reality: In this respect it is phenomenologically unique (a strictly singular entity); and (as we will show in Section 4) it can be conceived of as functionally unique by continuously being embedded into *itself*. Phenomenal holism not only reappears on sub-global levels, but it also has a measurable, *quantitative* dimension: Perceptual objects, multimodal scenes, or the phenomenal self are examples of conscious contents clearly possessing the additional feature of a *gradient of coherence*. Therefore, what these contents represent to the organism is also the internal correlation-strength between the features integrated. As this integration is of a non-compositional and non-conceptual nature, it also generates the quality of wholeness sketched above. An important epistemic goal would be to describe the *embedding relations* holding between various classes of subglobal wholes, and to the highest-order whole as well. We will offer one such model in Section 4.

Let us once again point out the relation between GU and H: GU is a special case of H by being the *highest-order form* of holism by which a subject's phenomenology can be described; in addition, it can only be ascribed to the totality of conscious experience, i.e. to the *whole* phenomenology (and not to parts of it). In a single subject, this property is never multiply instantiated. This gives GU a unique status and explains why we list it as an extra constraint.

- (GDI) Phenomenological constraint #3: Most conscious subjects describe the experienced unity as well as the holism of their subjective world as characterized by *global diachronic identity*. That is, although the contents of consciousness permanently change, there is a characteristic global sense of *sameness across time*: The one, single world in which all my conscious experiences unfold is consistently experienced as the *same* world.
 - GDI may not be a necessary condition for the ascription of consciousness, because there seem to be rare phenomenological configurations, in which it is not satisfied (as in dissociative identity disorder [DID], severe cases of amnesia, disorientation, dementia or the non-lucid dream state).
 - Diachronic identity again reappears on the level of subglobal wholes. For example, we typically have a distinct sense of sameness across time for the phenomenal self and for many perceptual objects (“object permanence”), including specific scenes, situations, behavioral spaces, etc. For example, the phenomenal self would be an obvious candidate to locate the sense of the diachronic unity of the “stream of consciousness” as such. Here, we will ignore this issue.

- On the representationalist level of description, GDI says that conscious experience is unique in that multiple relations of *transtemporal identity* are simultaneously represented, typically on multiple levels of granularity. Here, the centrally relevant epistemic goal lies in providing a functionalist analysis of this, preferably one that is grounded in a mathematical model.
- (STC) Phenomenological constraint #4: In the domain of subjective time experience, most conscious subjects describe their experience at a time as characterized by *strong temporal continuity*. This means that the phenomenal Now – the current “moment of experience” – is described as having introspectively indiscernible boundaries, i.e. moments of experience, although highly salient features of conscious experience, are not fully individuated phenomenal entities, they are underdetermined with regard to their beginning and their end (a very similar idea has been formulated as the notion of “extreme continuity”²). Of course, we have a subjective sense that they arise and pass away, come, perhaps linger a while, and so on – but there are no introspectively determinable boundaries, in principle.
 - STC currently is another candidate for a necessary condition for the ascription of consciousness, because (to the best of our knowledge) there seem to be no phenomenological configurations, in which it is not satisfied.
- (GC) Phenomenological constraint #5: Most conscious subjects describe the totality of their experiences as being *continuously* stretched out in time (*global continuity*; see also Section 4). If a globally unified state has emerged (GU), then it is also characterized by strong temporal continuity (STC). Global continuity has two aspects:
 - First, conscious experience is described as *temporally* integrated and globally coherent (“gap-free”);
 - Second, the transition from one phenomenal Now to the next is not introspectively accessible. Thus, GC picks out two senses in which conscious experiences can be described as diachronically connected. The first is, we claim, a necessary condition for the ascription of a phenomenal Now (which we will henceforth also call *temporally minimal experience*).

2. This term is borrowed from Oliver Rashbrook who defines extreme continuity as follows: “Consciousness exhibits extreme continuity if and only if the temporal boundaries of experience fail to be manifest in consciousness.” (2011:7). William James expresses a related idea when he writes that the specious present “[...] has, in addition, a vaguely vanishing backward and forward fringe [...]” (James 1890:613).

We have now confined the domain of inquiry by introducing a set of phenomenological constraints. What we are interested in – our overarching epistemic goal – therefore could be described as

- the conscious presence of a single, “dynamically nested” world,
- which has parts that themselves can have parts,
- and are experienced as forming units of varied internal connection strengths,
- that can change continuously in time without losing their numerical identity as internally represented.

But what are “phenomenological constraints” in the first place, given that no such thing as “first-person data” in a more rigorous sense exists?³ A phenomenological constraint is a soft constraint for theory-formation, emerging from the attempt to take subjects’ reports about what they describe as “their own conscious experience” seriously, namely, by treating such autophenomenological reports as data, extracting formal invariants from them and asking: What kind of internal dynamics would necessarily generate such kinds of reports in any given system? So our second goal in this chapter is to take classical heterophenomenology (Dennett 1991, 2003, 2007) one step further and find out if a specific set of conceptual tools – namely those provided by mereotopology (see Casati & Varzi 1999 for an introduction) – can help in extracting such formal invariants.

3. Philosophically assuming the existence of ‘first-person data’ (i.e. while operating on a *metatheoretical* level of inquiry) rests on an extended usage of a concept that is only well defined in another context of application (namely, first-order *scientific* theory formation targeting facts, not, say, the logic of sentences describing or predicting such facts). To begin, the whole concept of a ‘first-person perspective’ is just a visuo-grammatical metaphor, without a theory to back it up – currently, we simply don’t know what that could be, ‘a’ first-person perspective (for a first conceptual differentiation, see Blanke & Metzinger 2009). Second, ‘data’ are extracted from the physical world by *technical* measuring devices, in a *public procedure*, which is well-defined and well-understood, replicable, and improvable; and which is necessarily *intersubjective*. ‘Data’ are typically (though not always) gathered with the help of technical measuring devices (and not individual brains) and by groups of people who mutually control and criticize each other’s methods of data gathering (namely, by large scientific communities). The epistemic subject is a cooperating group of rational agents, not a subject of phenomenal states. In particular, data are gathered in the context of rational theories aiming at ever better predictions, theories that – as opposed to phenomenological reports – are capable of falsification. Autophenomenological reports themselves can be treated as data, the experience itself cannot (Dennett 1991, 2003, 2007). All of this is not to deny that what are sometimes called ‘first-person methods’ possess indisputable heuristic value and could have an enormous impact on our way towards a rigorous, empirically based theory of consciousness.

2. Mereotopology as a new tool for the study of consciousness

In this section we highlight the different ways in which our five constraints involve mereological notions (like *parthood*, *overlap*, etc.) and topological notions (like *connectedness*, *boundary*, *open*, *closed*, etc.). These descriptions must, at this point, remain rather general, but they already open paths to specific definitions that are more directly amenable to empirical application. Here, we proceed by highlighting open questions the answers to which can help in deciding which paths a formal mereotopological theory of (the unity of) consciousness ought to take.

- Constraint #1: As mentioned above, the degree of global holism displayed by an experience corresponds to the extent to which its parts are *whole* parts. A whole part is, as a first approximation, a part that is *connected* to other parts of the experience. This connectedness could, for instance, be described as a kind of causal coherence (on the functional level of description) or as a kind of correlation between random variables (on the informational-computational level) or as (synchronous) firing of groups of neurons (on the neurobiological level). In order to give accurate descriptions on all of these levels, it remains to be determined how best to characterize the parts of an experience in the first place:
 - Is it useful to define parts in an extensional way by listing classes of experiences that have these parts in common?
 - More specifically, could it be useful to define parts using a similarity relation (for instance by identifying a part with a class of experiences that are similar to each other to a certain extent)?
 - Or should any definition of phenomenal parts make reference to “phenomenal properties” (as, for instance, proposed by Timothy Bayne 2010:24)?
- Constraint #2: A subglobal whole is a whole part of an experience or a part of an experience; it is subglobally unified to a greater extent than the experience (or part) it is part of. An experience satisfies *H* if it has subglobal wholes and satisfies *GU*.
- Constraint #3: Global diachronic identity is a relation between certain temporal parts of “streams of consciousness”, viz. between phenomenal Nows. Such phenomenal moments must at least be experienced as being similar. But different temporal parts are never experienced at the same time. Therefore it seems that what is essential is that the temporal parts themselves have synchronic parts that are somehow similar to other temporal parts. That way, a phenomenal Now comes to be “contained” in the following phenomenal

Now, as it were. We propose the term *global self-embedding* for this process. Below (in Section 4) we provide an illustration of this idea in representational terms.

- Constraint #4: As a first approximation, two different phenomenal Nows can be described as displaying STC if they overlap.⁴ On the one hand, this indeed seems to capture the idea that moments of experience are underdetermined with regard to their beginning and end, because the overlapping temporal part is the end of the temporally earlier and also the beginning of the later moment of experience. On the other hand, we will argue below that moments of experience can only be said to have temporal parts if moments of experience can be *identified* with neural processes. Hence overlap between phenomenal Nows cannot be used to analyze STC without making further assumptions.

A different possibility can be suggested if global self-embedding can indeed be analyzed as a process which is constituted by processes of local meta-representations (as we suggest below); in that case, if the local processes are not completely synchronous, the global process will have vague temporal boundaries.

- Constraint #5: The second aspect of GC implies that experiences are not experienced as changing abruptly. This makes it so tempting to speak of a “stream of consciousness” that we as subjects of experience observe, like a river flowing continuously before our eyes. Upon reflection, however, we recognize that we are actually part of the river ourselves (cf. Metzinger 2004:153): our phenomenal self-models are parts of the overall experiences. That is why we cannot gain introspective access to what happens “in between” two phenomenal moments. But neither is this possible with respect to *parts* of an experience. Why is this so? If, on the representationalist level of description, consciousness necessarily involves meta-representations, then there is the possibility that these meta-representations represent some of the synchronic and diachronic features of their intentional objects in a coarse-grained way, thereby integrating them to more coherent and more abstract “chunks”. In other words, a meta-representation of a representation R might have the intentional object “R represents O at time t”, but it might also have an intentional object like “R represents O over a certain period of time during which this or that feature remains the same and some other feature changes”.⁵

4. Invoking the notion of overlap to explain the diachronic unity of consciousness is not a new idea. Recently it has e.g. been advocated by Barry Dainton, see Dainton (2010).

5. An idea which comes close to this has been employed by Jun Namikawa and Jun Tani in a model in which subsystems at different levels learn to predict features at varied levels of

It remains to be seen if this interesting kind of meta-representation could also be realized by a self-embedding process.

Accepting that our constraints involve mereological and topological notions, one might still wonder why formal mereotopology should be useful for their further analysis. There are at least two reasons why this can be fruitful: On the one hand, by considering whether mereotopological axioms and theorems might be applicable to the domain of conscious experience, one can gain a completely new perspective on the problem of consciousness while simultaneously generating new, and more precise philosophical and empirical questions; on the other hand – and this is the main point – we see mereotopology as a tool for *integrating* constraints on different levels of description. Why is this important? The MICS methodology (see Section 1) not only involves posing constraints, but also studying if and to what extent descriptions on different levels actually pick out the *same* features. Therefore, if one wants to integrate different theoretical approaches (or existing data-sets), it is vital to create formal bridging principles that reliably connect different levels of analysis.

Let us illustrate this with an example: The idea that conscious systems are organized in such a way that they are both functionally segregated and integrated has become quite popular in consciousness research. This can be seen by the growing number of formal methods which are being proposed as ways to measure the extent to which a system exhibits this kind of dynamical complexity (see for instance Tononi & Edelman 1998; Tononi & Sporns 2003; Tononi 2008; Barrett & Seth 2011; and Seth, Barrett & Barnett 2011). All these measures seem to capture something similar (viz. the extent to which a system is both *integrated* and *differentiated*), but what is lacking is a conceptual clarification of integrated information that is general enough to cover different ways of operationalizing this kind of dynamical complexity. This may on the one hand seem premature, because most measures are only of limited value in practical applications anyway, whence the need for a unifying framework may be thought not to be very pressing; on the other hand, as integrated information has been claimed to be *identical* to consciousness (see Tononi 2008), it becomes crucial to specify those aspects of consciousness that make it “integrated and differentiated” and say why some particular formal concepts can serve to explain these aspects. A formal mereotopological theory can help with this task by being general enough to be applicable to different domains

abstraction: “The higher level learns not for details of sensory–motor patterns but for abstraction of primitive sequences with long-term dependency.” (Namikawa & Tani 2008: 1467). Note however, that this can arguably only be regarded as an example of re-representation, not of meta-representation (let alone self-representation).

and thereby also providing hints as to what a general account of integrated information could look like.

Furthermore, while it is clear that phenomenal experience has a mereological deep structure (see Constraint #4 in Metzinger 2003, Section 3.2.4); there is no explicit account of the internal mereotopological structure of conscious experiences. On the other hand, the idea that conscious states can be described as points (or trajectories) in a ‘phenomenal state space’ that has a certain topological (or even metric) structure is shared among many researchers.⁶ This structure is induced by what Tomer Fekete calls *primal similarity* (see Fekete 2010:75–76), i.e. the fact that different phenomenal experiences can be grouped according to whether they are more similar or not: a pain in the thumb is – all other things being equal – more similar to a pain in the index finger than to the feeling of swimming in a cold lake. This means that we can demarcate clusters in phenomenal state space in which similar, though distinct, experiences reside. Such clusters can, in principle, give rise to arbitrarily complex topological structures, and this should also be reflected in whatever activity gives rise to conscious experience, if we assume that a structure-preserving function⁷ exists between phenomenal state space and neural state space (or, more generally, activity space – that’s the term Fekete uses). Fekete has also proposed a formal way of measuring the topological complexity (*representational capacity*) of a given activity space (see Fekete 2010 for a description of the approach, Fekete et al. 2009 for an application, and Fekete & Edelman 2011 for a more general motivation of the approach).

But there is not only an interesting internal structure to phenomenal state-space, but also to individual phenomenal states themselves. They possess internal depth. At first sight, it seems that measures of integrated information only capture the structural complexity *within* phenomenal states. Fekete’s representational capacity, on the other hand, explicitly takes both the complexity of trajectories in activity space and the complexity of the space constituted by these trajectories into consideration. The idea now is that a general mereotopology for consciousness could help clarify the way in which these approaches are interrelated and, additionally, help integrate them with purely philosophical approaches. Recall that our overall goal in this section has been to sketch a route towards constraint

6. See Fell (2004) for an overview (which traces the idea that there is a homeomorphism between phenomenal state space and neural state space back to Müller 1896).

7. By “structure-preserving function”, we mean at least a homeomorphism. If phenomenal state space and activity space can be modeled as metric spaces it will be reasonable to expect an isometric isomorphism (as Fekete 2010 does).

integration on a more formal, fundamental level. The next section can be seen as a critical case-study in the same spirit.

3. A critical application: Does an experience have temporal parts?

Does it make sense to speak of the “duration” of an experience, or, conversely, of it only occupying a “point in time” (cf. Dainton 2010)? We do not think so. At least on a representationalist level of analysis, experiences are the *contents* of ongoing processes of phenomenal representation. A representation is something that can be individuated via its content, its carrier, and its format. Let us begin with the last characteristic feature. A “format” is an *abstract* property having to do with how information is encoded and decoded, and with what mechanisms can access it – just think of propositional vs. pictorial formats, symbol chains vs. vector transformations, conceptual thought as opposed to mental imagery. A format is an abstract property of representations or data-structures, it is the form or **mode of presentation**, and we may helpfully view it as a functional property realized in an information-processing system, allowing raw data to be interpreted, processed, saved, reopened, or integrated in a highly specific manner. “Phenomenological formats” then become abstract properties of conscious representations, determining a specific inner phenomenal mode of presentation, realizing specific functional properties – examples would be visual format, auditory format, somatosensory or kinesthetic format etc. The “carrier” of a phenomenal representation is the concrete, underlying neural dynamics in the physical world, and, of course, here it makes perfect sense to speak of temporal properties like “duration”, “minimal temporal extension”, etc. However, the experience itself is another abstract property of this physical process, namely the “content” it has. Of course, it is possible to have a subjective experience *of* duration, *of* continuity, or *of* minimal temporal extension or “Nowness”, etc. But as long as we stay with the representationalist level of description, the content constituting the conscious experience itself is an abstract entity devoid of temporal properties. Therefore, it makes no sense to ask or discuss what the “duration of an experience” could be.⁸ Obviously, one can always compare apples and oranges by measuring the physical “clock time” of a given external event, and comparing this to the estimate of this duration which a conscious person can produce while solely relying on her phenomenal representation

8. Cf., however, the discussion in the following paragraph, where we deal with the question if the temporal properties of representational carriers could be used to derive the “duration of an experience”.

of this event.⁹ But this confuses a physical property of the carrier with an abstract phenomenal content property (the second dimension of individuation with the third, if you will). There is, it seems, only one way to speak of the “duration” (or any other temporal feature) of an experience, and that is by making a classical identity claim and eliminating the representational level of description altogether: If we can show that what we today call conscious “contents” is strictly identical with what we have called the “carrier” above, that it shares all non-intensional and non-modal properties of the underlying neural dynamics, *then* we can speak of an experience having, for example, a duration. The duration of the minimally sufficient neural process would then be the duration of the experience.

The following might be objected to this: suppose we have a temporal chain of representational carriers that form a continuous process underlying the experience of watching a movie (see Figure 1). This process can be divided into temporal parts which themselves constitute carriers of experiences (say, of the experience of watching the first 10 minutes of the movie, the experience of watching the next 10 minutes, etc.; see Figure 2).

Hence, the experience of watching the movie can be divided into parts that are the contents of *temporal* parts of the process underlying the whole experience.

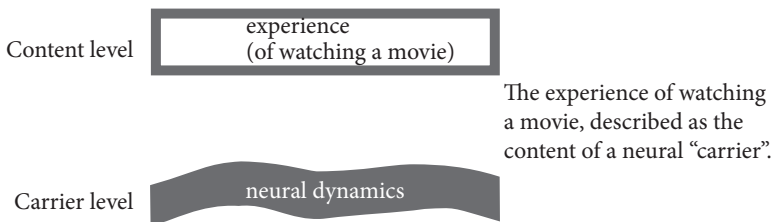


Figure 1. Temporally continuous representational carriers.

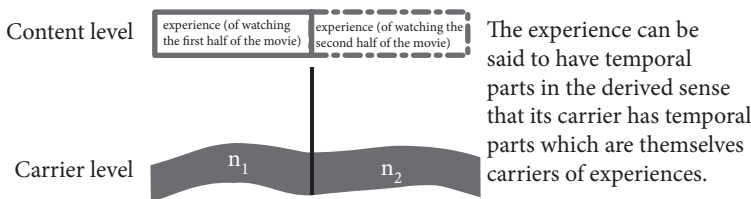


Figure 2. Experiential parthood as conceptually derived from partitioning the carrier.

9. In this case we would compare the clock time of an event to the estimate of the duration of the *event* experienced, but not to an estimate of the duration of the *experience* of the event.

And, it might be added, it is in this sense that the experience (if conceived of as the content of a representation) can be said to have temporal parts. Hence, although we cannot attribute objective temporal properties to contents in a strict sense, there is a derived sense in which contents can be said to have duration, temporal parts, etc. Is this convincing? Of course, no one can be barred from calling something a temporal part of a content. For us, however, the question is if this way of talking is *explanatorily useful*.

We noted above that such talk of temporal parts of experiences might be useful to explain the strong temporal continuity (STC) of consciousness: If two experiences overlap in time, the overlapping part is both the “closing part” of the first experience and the “opening part” of the second experience. Therefore no one will ever experience a boundary of a phenomenal Now, because it would at the same time always have to be experienced as the boundary of another phenomenal Now (so it cannot be experienced as either the beginning or the end of an experience). What we *really* experience is the end of a beginning and the beginning of an end. Appealing as this might perhaps sound, let us observe what this proposal comes down to if we consider the *strict* sense in which there are overlapping parts (see Figure 3).

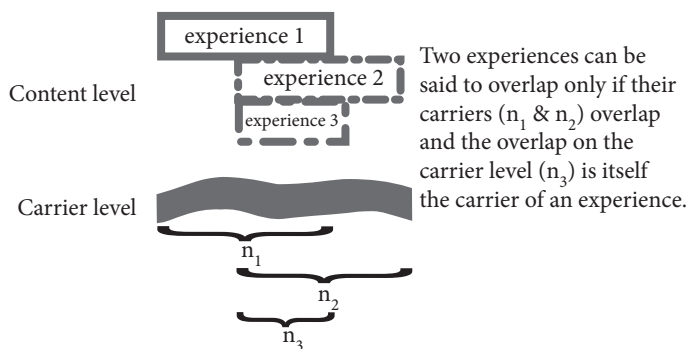


Figure 3. Experiential parthood as derived from a rigorous definition of “overlapping”.

That two experiences overlap means, strictly speaking, that their underlying “carriers” are processes which overlap in such a way that the overlapping part is itself the “carrier” of an experience. But this experience will display STC, as well. How could we explain this formal feature? According to the above proposal, it overlaps with another experience. The overlapping part must again be an experience, and so on. The problem with this is not that we are running into an infinite regress, but that – given our own empirical constraints – we can only make a *finite* regress. For we take as a background assumption that the process which underlies an experience is temporally extended and that there is a lower bound

greater than zero on the time spans it takes for such a process to unfold.¹⁰ Hence, any experience will have a finite number of phenomenal temporal parts (in the sense explained above). Therefore, we can (in principle) divide an experience into temporally minimal experiences,¹¹ experiences that do not have temporal parts that are themselves experiences (which means, strictly speaking, that we can divide the “carrier” of any experience into shorter “carriers” such that these do not have temporal parts which are themselves carriers of experiences). Consequently, minimal experiences do not overlap (not even in the loose sense), so overlap cannot explain STC – unless, once again, we can identify experiences with the neural processes giving rise to them.

4. “Self-embedding”: Sketch of a positive proposal

In this section we will propose a positive model for an embedding relation (see Section 2), with a focus on the global variant. In order to illustrate the idea of a global self-embedding process, we provide an interpretation in terms of *self-representation*, where a self-representation is a representation which also meta-represents itself. That is, it is not a representation of some mysterious object – “the” self – but a process that continuously monitors itself, albeit in a very specific way. One aspect of this formal model is that it describes the emergence of an *internal* meta-representation, a second one is that it explicitly introduces the *temporal* dimension, yet another one is that the specific combination of integration and global meta-representation (which we take to be the two core characteristics of conscious experience; cf. Metzinger 1995) is here conceived of as a dynamic bottom-up process exhibiting *indirect semantic coupling* on a subsymbolic level of representation. This interpretation is inspired by a proposal by Kenneth Williford, according to which “[...] consciousness is the product of the coupling of two representational systems, and [...] the self-representation characteristic of consciousness is achieved indirectly via that coupling.” (2006: 134). We will first extend this idea by providing a description in terms of forward models – which thereby also suggests a possible answer to the question *to what end* two representational systems should come to be coupled (in the way specified below) in the first place.

10. The assumption that not every temporal part of the “carrier” of an experience is itself the carrier of an experience is argued for in Phillips (forthcoming), see also the classic discussion in Dennett & Kinsbourne (1992).

11. In this paragraph, we will also use the term *minimal experience* as shorthand for *temporally minimal experience*.

That is, we want to explicitly add the temporal dimension and ask for an empirically plausible model: In an evolutionary context, what could have proved to be a *functionally adequate* architecture?

In what way could a representational system be embedded into itself? Think of a maximally simple system first, comprising only two parts, R1 and R2, which each serve to represent some aspects of an object O. It is suggestive to conceive of the two subsystems as implementing forward models which emulate how certain efferent signals will influence the feedback signals that will be obtained by a “measurement” of the object O (cf. Grush 2004; Franklin & Wolpert 2011). In addition to the forward models that stand in for aspects of the represented object, the two systems R1 and R2 each also comprise a forward model that emulates certain aspects of the other system. In effect, R1 emulates the behavior of O and also the behavior of R2. We are strictly speaking about subpersonal processes here, but let us introduce a personal-level metaphor to illustrate our first point: R1 “knows” that R2 also looks at O. Likewise, R2 emulates O (possibly with an “emphasis” on different aspects of the object), and R2 also emulates R1. The intentional object of R2 is “O as currently related to/represented by R1”, whereas the target of R1 is “O as currently related to/represented by R2”. Note that by emulating R1, R2 emulates a system that emulates R2, so R2 emulates itself in an indirect way (the same holds for R1). In order to learn how to emulate the other system, each of the two coupled systems must receive some feedback from the other system, thereby obtaining a way to evaluate the quality of its predictions.

It is crucial to explore possible merits of such a pair of representational systems, because the added “burden” of representing the other system might at first sight seem completely dispensable or functionally inadequate. Clearly, in real-world biological systems a metabolic price would have to be paid for such an extension of the organism’s representational architecture. What could justify paying this price? Here, we can only suggest two highly speculative possibilities.

- Continuously emulating the behavior of another representational subsystem might increase the *quality* of the individual representations – internally generated information becomes more *reliable*.
- Evaluating how well the behavior of another representational system is predicted might serve to examine how well the representations fit with each other. This would provide a way of increasing the *coherence* of currently active representations via enhanced functional interdependence.

We could extend our personal-level metaphor again by imagining two soccer players trying to gain control of the same ball, while mutually looking at each other in order to find out what the other player will probably do next. Think of a system in which different subsystems compete for access to some higher processing

level (e.g. in the spirit of the model described in Namikawa & Tani 2008). This higher processing level could, for instance, consist of some suitable, neurobiologically constrained version of a *global workspace* (cf. Baars 2007). The winner of the competition is the subsystem that is able to predict some feedback most accurately. In such a competition, being able to predict the behavior of the competitors to some extent could turn out to be of advantage. If each system receives more than just the feedback it is trying to predict, but also feedback indicating how well some of the predictions of its competitors were, the systems could “decide” to incorporate a prediction of the behavior of another system into its predictions of the future behavior of their commonly shared target object.¹² Which of the competitors is being emulated could also be dependent on the context. In this way, every subsystem would have to “copy” the behavior of some of their successful competitors to some extent, thereby improving its own performance. This would be a marked extension of the idea that the brain employs (hierarchical) predictive coding (by introducing lateral predictions, as it were), but as the idea is so widespread and used to explain such a broad variety of cognitive capacities (see e.g. Kilner, Friston & Frith 2007 and Seth, Suzuki & Critchley 2012), we think that any extension of it is at least worth considering.

Second, when two competitors are both successful in predicting the behavior of some object (possibly by emulating slightly different aspects of its behavior), they will necessarily incorporate emulations of each other in their predictions (if they follow the strategy described above), and this will arguably make all currently active representations slightly more similar and also globally coherent. In terms of our personal-level soccer analogy: A “game” emerges. The coherence of the representations could be assessed without “comparing” them directly, but just by evaluating whether both systems are *successful* in emulating the respective other system: If R1 generates a good prediction and if this prediction is partly arrived at by using a (good) prediction of R2, and if the same holds for R2, then the predictions by R1 and R2 will be coherent; because they have both been generated by incorporating (a good prediction of) the other one. Hence (and this is the central aspect of the idea we propose), R1 and R2 can be said to fuse into a *single* representational system R_g (at least temporarily), partly because a new and *global* dynamic property has emerged.

12. In a more general vein, Malach (this volume) argues that in assemblies of neurons, information about the state of the whole assembly can be made available to each of the neurons in that assembly, thereby driving the activity of the assembly to reach a fixed point (or to come close to a fixed point attractor). A mechanism like this could turn out to be necessary to implement our proposed GSE architecture.

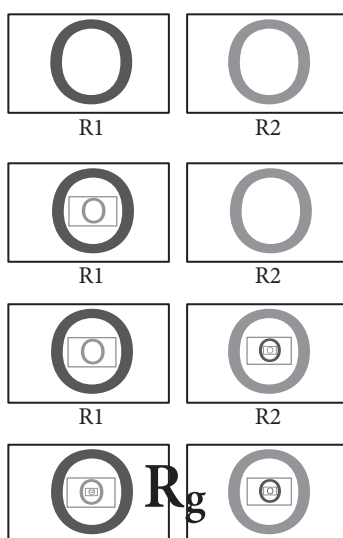


Figure 4. The GSE-architecture: Global self-representation and integration as achieved by local and mutually coupled forward models. The figure illustrates four steps which can lead to a functional coupling between two representational subsystems, such that a new global self-representation (R_g) emerges. Rectangles stand for representational vehicles, everything inside the rectangles stands for representational content. The letter “O” indicates that R1 and R2 have the same *target* object, while their *intentional* objects can partly differ (i.e. the properties “ascribed” to O by R1 and R2 can be slightly different; this is indicated by the two different shades of grey).

Here is a short description of the four steps illustrated by this figure:

1. R1 and R2 both represent O.
2. R2 represents O; R1 represents O and R2.¹³
3. R2 represents O and R1 (which represents R2; therefore R2 indirectly represents itself). R1 represents O and R2.
4. Both R1 and R2 represent the respective other representation and O. Therefore, the conjunction of their representational vehicles can be described as the vehicle of a new global meta-representation R_g .

13. Please note how this intermediate step – which makes for an asymmetry between R1 and R2 – is not strictly necessary. If, however, representational subsystems only start representing other representational subsystems which generate successful predictions, then such an asymmetry can arise, for instance, if there is only one successful representational subsystem at a certain time (which therefore will not start emulating any of the other systems at that point in time).

The philosophical idea behind this architecture can be formulated as the “Principle of Globalization by Representational Fusion”: (PGRF):

- If two representational systems previously treated as distinct functional entities track (different features/parts of) the same target object¹⁴ over a certain period of time,
- and if in doing so they are functionally coupled in an appropriate manner,
- then, during this period of time, we may speak of a newly emerged representational carrier *as well* as of a newly generated representational content.

PGRF is a conceptual principle. Here, we have done only two things: We have offered one concrete application of PGRF in the temporal domain; and proposed an architecture that spells out what an “appropriate” functional coupling could be. Of course we can also imagine larger multi-player systems with more than two representational systems mutually tracking each other. Other representational systems will tend to incorporate predictions of their (joint) behavior in their own predictions; this could in principle lead to further couplings, say, between a system R3 and R1 & R2 (provided the predictions of R3 somehow complement the ones generated by R1 & R2). A coupling of more than two representational subsystems would require, strictly speaking, that each of the subsystems comprises at least one model of each of the other subsystems. This may seem utterly implausible. In order to avoid an overly complex architecture, we propose that the subsystems only emulate certain aspects of the other systems, and that the amount of detail thus modeled can change dynamically and need not be distributed evenly among the other subsystems (one could envision a

14. In the soccer analogy, this same target object could be the ball or the ball plus the position of the players or (parts of) the match (this would be a more abstract object). In individual conscious human beings, the common object targeted by different representational subsystems could be the body (with each subsystem being “specialized” in representing some part of the body), or it could consist of those parts of the world which are currently relevant for the system. In general, such an object should arguably satisfy the following two conditions in order for PGRF to be applicable: The object should have different parts the relevant properties of which are, on the one hand, correlated in such a way that representing those parts by separate subsystems is more efficient than using one central representation of the whole object (this would also guarantee that the representations would not be *locally* incoherent); on the other hand, the available information regarding each of the object’s parts should be underdetermined to the extent that a desirable degree of *global* coherence *cannot* be guaranteed. Our speculative hypothesis is that PGRF would in this case provide a way of generating a new global coherent representation *using a minimal amount of additional resources*. For an extended discussion with respect to the body as target object, see Alsmith (forthcoming).

process of “selective feature-binding” leading to a variable landscape of more or less intense local couplings).

Again the soccer metaphor can be of avail: in principle, it would be beneficial for every player to employ representations of the exact positions and intentions of every other player. In practice, however, only a fraction of this information is needed, depending on the situation the player is in. When trying to hinder an attacking player from breaking through the defense of her own team, a player does not need to know what is happening in the opposite half of the pitch; when she wants to make a long pass, however, the positions of players in the other half can be relevant. Similarly, every subsystem should be thought of as specializing on certain aspects of an object. Hence, only some aspects of the other systems’ outputs will be relevant, and this will depend on the situation. This is not to say that a system with multiple coupled subsystems must always be functionally adequate or even neurobiologically plausible (this would have to be answered by future research); the point of this paragraph is merely that it is not as implausible as it might seem at first glance. Furthermore, our model does not strictly require multiple couplings, but only at least one coupling (between two representational systems).

The philosophical point behind our proposed architecture is that

- it turns the overall representational system into a **self-representational system** via the described mechanism of dynamically coupled **local meta-representations**;
- such a system thereby necessarily creates a single **globally coherent state**;
- in the temporal domain, this newly emerged global state is a good candidate for what, in the past, we have often called the “**specious present**” or simply “the” phenomenal Now (see below);
- and it is a system which, on a microfunctional level, continuously **self-embeds its current state into the next global state**.

If this is correct, then it should also be clear that our system needs at least one time-step in order to achieve the required form of interlacing leading to a globally coherent state of self-representation. We have therefore offered at best a model of how to satisfy GU and H, but we need to turn to the diachronic constraints STC, GDI and GC to show that such a global state can indeed have properties that are often ascribed to the phenomenal Now.

In order to do so, one should investigate how the model could actually be implemented; the temporal properties of possible implementations would then provide suggestions on how to satisfy the diachronic constraints. One possibility regarding STC has been mentioned above: if the local meta-representations are not completely synchronous, the global self-representation will have vague temporal boundaries.

GC: Above we mentioned that the reason why the transition between phenomenal “Nows” seems to happen continuously might be due to temporal differences, respectively the functional level of “grain” employed by the system through which it internally individuates temporal entities, i.e. due to differences between the intentional objects targeted by representations and by meta-representations. The basic idea is the following: think of a conscious system that comprises representations that represent certain aspects of an object at particular points in time (or short periods of time). For instance, the intentional object of such a representation might be “object O at time (interval) t1”. In addition, there can be representations that represent certain aspects of an object over a longer period of time. The intentional object of such representations would be something like “object O at times t1, t2, and t3”. More specifically, the object could thus be represented as having certain features that stay the same during the time interval defined by t1 to t3, and as having other features that change during this interval. Such representations (which represent features abstracted from a longer period of time) could also be meta-representations. Their intentional objects would then be, for instance, “object O as represented by R at times t1, t2, and t3”. The interesting question we want to approach in this paragraph is whether a system could represent an object at different times and also meta-represent itself as representing the object at different time scales. In control-theoretic terms, the system would comprise forward models that produce estimates of an object at different times, and in addition it would also produce estimates of successions of such estimates.¹⁵

Let us start again with a maximally simple model consisting of two representational subsystems R1 and R2. Similarly as before, each system predicts at each “time step” some feedback of an object O. In addition, both subsystems generate estimates of the behavior of the other representational subsystem over multiple time steps. So R1 will at each time step represent O and also O as represented by R2 at multiple times. The same holds, *mutatis mutandis*, for R2. Hence, each representational system will (indirectly) represent itself as representing O over multiple time steps.¹⁶ That way, R1 could represent features of O

15. In a way, this would be a meta-representational version of Rick Grush’s *trajectory estimation model* (see Grush 2008).

16. We have not said much about why a representational system R1 that represents another system R2 (that in turn represents R1) should be said to represent itself (thanks to Kenneth Williford for pointing this out). This is a problem which deserves a broader discussion than we can provide in this chapter. As a beginning, if representational subsystems emulate each other to “learn” making good predictions, then a system R1 may gain an additional benefit from representing another system R2 by trying to find out what that other system “likes”

and also (indirectly) represent itself as representing features of O as changing or as staying the same. R1 would thus not only represent features of O at, say, time t1, but it would also represent features of O at other times, say t0, t1, and t2 – by meta-representing itself (indirectly) as representing those features at t0, t1, and t2. According to this idea, the transition from one phenomenal Now to the next is in principle inaccessible, because no phenomenal Now is ever experienced in isolation; instead, the temporal context (estimates of the foregoing and the following phenomenal Now) is always embedded in any given phenomenal moment. There still is a context, but it is not conceived of as an external, rigid frame of reference any more. It now has itself become an entity that continuously evolves over time and forms an *internalized* set of functional boundary conditions for the overall process.

GDI: This speculative proposal might even suggest how to satisfy **GDI**. In order to be represented as being identical in a certain sense, there must be a sense in which successive phenomenal moments are represented as staying the same. We noted in the previous paragraph that one can conceive of representations that comprise representations of themselves at different times, and that these meta-representations can be more or less abstract, for instance by having an intentional object like “R1’s representation of O changes only marginally during t1, t2, and t3”. A system having a GSE architecture could thus also *globally* represent itself as having certain invariant *global* features (i.e. features that must now be ascribed to the system as a whole, and not to any of its subsystems any more), and this might be a way to satisfy **GDI**. To state it more clearly: if a representational system can represent itself as having certain stable features (and if these features can be ascribed to the system as a whole, i.e. not only to parts of the system), then this might play a crucial role in explaining why consciousness is described as displaying global diachronic identity. A related way of thinking about this is that one could conceive of global diachronic identity as ultimately being derived from the diachronic identity of objects (as constituting parts of a “phenomenal field”). The idea would be that we experience the world as being the same world, because we experience most of its parts as persisting through time. Of course, not all parts of the world are actually experienced as persisting, but

about R1. In other words, R1 could try to evaluate in what way the fact that R2 represents R1 enhances the quality of R2, and this could provide another way of improving its own predictions (apart from any external feedback it gets) or it could simply serve to increase the overall coherence of the representations. Regarding this issue it seems pivotal to us to specify explicitly how one system R1 could determine in what way its own activity influences the activity of another system R2.

there are some parts which are there all the time and which are almost permanently consciously experienced by us: our body parts (cf. Metzinger 2004: 291). If representations of different body parts are coupled in the way indicated above, a new global meta-representation could represent its representational parts (i.e. the representations of body parts) as jointly representing something more abstract and global, viz. *the* body, namely, as a single functional entity. This more abstract intentional object can remain the same even when its parts change (i.e. we experience our body as being the same body over time – in pathological cases the experience can of course change without there being a significant change in the body parts). So far, this is only a representation of an object that remains the same (or an abstract representational content that stays the same, if you will). If we add the temporal dimension to our description, however, we see that the global meta-representation can not only represent its parts as jointly representing an object (which in fact is invariant), but it can also represent them as jointly representing this object *as being* invariant. This is achieved by abstracting from particular varying features of the body and by coarse-graining the timescale at which the representational subsystems operate. If this is convincing, then there is no reason to doubt that a conscious system could similarly represent itself as representing certain abstract features of the *world* as being invariant. And this could give rise to the experience of a world that remains the same.

Again we must emphasize that these proposals on how to satisfy our constraints by a model serve primarily the purpose of illustration. A rigorous treatment would have to be based on a fully formalized version of the GSE-model. At the same time, such a model would only form one possibility, and it could even turn out that certain of its features would make it completely implausible as a model of how the brain *de facto* generates unified consciousness. This brings out the merits of a mereotopological approach: by first exploring logically consistent ways of analyzing the relations between the mereological and topological notions used in the re-formulated constraints in Section 2, one can develop more precise descriptions which are still general enough to be compatible with different approaches. This can provide a way to create various models that need not follow the path indicated by the GSE model. But still the GSE model can serve to get the mereotopological analysis started: the different kinds of representation involved in the model can, for instance, be regarded as realizing different ways in which subsystems can be connected to each other. “Connectedness” is clearly a more general term, so formal descriptions of ways of being connected could also be applied to non-representational systems (or to non-representational descriptions of systems). What is more, a formal analysis of “connectedness” could also suggest other ways of realizing global self-embedding by representational systems.

5. Conclusion

We have introduced a set of phenomenological constraints, trying to capture one centrally important aspect of phenomenal experience more precisely: the conscious presence of a single, “dynamically nested” world, which has parts that themselves can have parts, and are experienced as forming units of varied internal connection strengths, that can change continuously in time without losing their numerical identity as internally represented. We have then used mereological and topological notions to reformulate these constraints and offered two case-studies, in order to demonstrate the relevance and the viability of the new approach we propose. The first case study (Section 3) has been mostly critical, discussing the question of what it could mean to say of an experience that it possesses “duration” and “temporal parts”. The second case study (Section 4) was more constructive and included a positive proposal, offering one possible architecture (the “GSE-architecture”), with the philosophical idea behind this architecture (the “PGRF-principle”) being the notion of dynamically coupled, but localized processes of mutual meta-representation which can, over time and via a process of representational fusion, lead to a new global state which can in turn be characterized as possessing a new, global form of content. This content is a good candidate for what has been called the “specious present” or the “phenomenal Now”. We have called this process “self-embedding”. The general aim of this chapter has been to make a very first beginning in exploring the feasibility of mereological and topological notions. This was done in the context of the general methodology of multi-disciplinary constraint satisfaction (see Section 2), as a potential route to more precise descriptions of conscious processing, which have the capacity to *integrate* different sets of constraints and research programs operating on different levels of description.

References

- Alsmith, A.J.T. (forthcoming). *The Structuring Body: A Critical Study in the Description and Explanation of Perceptual Experience*. Mainz: Unpublished doctoral dissertation.
- Baars, B.J. (2007). The global workspace theory of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell Companion to Consciousness*. Malden, MA: Blackwell Publishing. doi: 10.1002/9780470751466.ch19.
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13(1), 7–13.
- Barrett, A.B. & Seth, A.K. (2011). Practical measures of integrated information for time series data. *PLoS Computational Biology*, 7(1), e1001052. doi:10.1371/journal.pcbi.1001052.
- Bayne, T. (2010). *The Unity of Consciousness*. Oxford: Oxford University Press.

- Casati, R. & Varzi, A. C. (1999). *Parts and Places. The Structures of Spatial Representation*. Cambridge, MA: The MIT Press.
- Dainton, B. (2010). Temporal Consciousness. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved February 24, 2012, from <http://plato.stanford.edu/archives/fall2010/entries/consciousness-temporal/>
- Dennett, D.C. (1991). *Consciousness Explained*. Boston: Little, Brown & Co.
- Dennett, D.C. (2003). Who's on First? Heterophenomenology explained. *Journal of Consciousness Studies*, 10, 19–30.
- Dennett, D.C. (2007). Heterophenomenology reconsidered. *Phenomenology and the Cognitive Sciences*, 6, 247–270.
- Fekete, T. (2010). Representational Systems. *Minds & Machines*, 20, 69–101. doi:10.1007/s11023-009-9166-2.
- Fekete, T. & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*, 20(3), 807–827. doi:10.1016/j.concog.2011.02.010.
- Fekete, T., Pitowsky, I., Grinvald, A. & Omer, D.B. (2009). Arousal increases the representational capacity of cortical tissue. *Journal of Computational Neuroscience*, 27, 211–227. doi:10.1007/s10827-009-0138-6.
- Fell, J. (2004). Identifying neural correlates of consciousness: The state space approach. *Consciousness and Cognition*, 13(4), 709–729.
- Franklin, D.W. & Wolpert, D.M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, 72, 425–442.
- Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27, 377–442.
- Grush, R. (2008). Temporal representation and dynamics. *New Ideas in Psychology*, 26 (2), 146–157. doi: 10.1016/j.newideapsych.2007.07.017.
- Kilner, J.M., Friston, K.J. & Frith, C.D. (2007). Predictive coding: An account of the mirror neuron system. *Cogn Process*, 8(3), 159–166.
- James, W. (1890). *The Principles of Psychology*. New York: Henry Holt.
- Metzinger, T. (1995). Faster than thought. In T. Metzinger (Ed.), *Conscious Experience*. Thorverton, UK: Imprint Academic.
- Metzinger, T. (2004). *Being No One. The Self-Model Theory of Subjectivity*. Cambridge, MA: The MIT Press.
- Mroczo-Wąsowicz, A. (forthcoming). *Die Einheit des Bewusstseins und das Phänomen der Synästhesie*. Mainz: Unpublished doctoral dissertation.
- Müller, G.E. (1896). Zur Psychophysik der Gesichtsempfindungen. *Zeitschrift für Psychologie*, 10, 1–82.
- Namikawa, J. & Tani, J. (2008). A model for learning to segment temporal sequences, utilizing a mixture of RNN experts together with adaptive variance. *Neural Networks*, 21, 1466–1475.
- Phillips, I. (forthcoming). The temporal structure of experience. In D. Lloyd & V. Arstila (Eds.), *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*. Cambridge, MA: The MIT Press.
- Rashbrook, O. (2011). The continuity of consciousness. *European Journal of Philosophy*, 1–30. doi:10.1111/j.1468-0378.2011.00465.x.
- Seth, A.K., Barrett, A.B. & Barnett, L. (2011). Causal density and integrated information as measures of conscious level. *Phil Trans R. Soc. A.*, 369, 3748–3767.
- Seth, A.K., Suzuki, K. & Critchley, H.D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2. doi:10.3389/fpsyg.2011.00395.

- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biol. Bull.*, 215, 216–242.
- Tononi, G. & Edelman, G. (1998). Consciousness and complexity. *Science*, 282(5395), 1846–1851.
- Weisberg, J. (2006). Consciousness constrained: A commentary on *Being No One*. *PSYCHE*, 11(5).
- Williford, K. (2006). The self-representational structure of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-Representational Approaches to Consciousness*. Cambridge, MA: The MIT Press.

The brain and its states

Richard Brown

LaGuardia Community College, NY, USA

1. Introduction

In recent times we have seen an explosion in the amount of attention paid to the conscious brain from scientists and philosophers alike. One message that has emerged loud and clear from scientific work is that the brain is a dynamical system whose operations unfold in time. Any theory of consciousness that is going to be physically realistic must take account of the intrinsic nature of neurons and brain activity. This important idea is often taken to be in conflict with the more traditional way of thinking about the mind in terms of static states like beliefs, pains, or representations of blue. One of the aims of this chapter is to show that this is not the case. We must use the traditional theories as a way to pick out the brain activity that we are interested in. In this way the two ways of thinking depend on each other. To corrupt a common Kantian line: uninterpreted brain data is meaningless and interpretation without brain data is blind. Once we adopt this model of scientific explanation and reduction we can build a case that phenomenal consciousness may turn out to be nothing but patterns of synchronized neural activity in specific frequencies against a dynamically changing chemical background.

To give some of the overall structure of my argument, in the next section I distinguish between Creature, Transitive, State, Access, and Phenomenal consciousness. Creature consciousness could plausibly turn out to be nothing but the global chemical state of the brain (in particular the exact ratio of aminergic to cholinergic neuromodulators) and transitive consciousness to be nothing but synchronized neural activity in various frequencies. Once we have identified these two fundamental kinds of consciousness in the brain the other three can be reduced to transitive consciousness. To see how phenomenal consciousness could be reduced to a kind of transitive consciousness I introduce the Higher-Order Representation of a Representation (HOROR) theory of phenomenal consciousness, which is a variant of the traditional higher-order thought theory. The main difference consists in the claim by the HOROR theory that phenomenal consciousness just is a kind of representation, albeit a higher-order representation (of a suitable kind).

As such there is no relation between the higher-order representation and any other representation needed. This distinguishes the HOROR theory from at least some versions of the traditional higher-order thought theory.

The HOROR theory strikes many people as counter-intuitive in that many people wonder how it could possibly be the case that the conscious experience of a pain – the painfulness of the pain – could be a higher-order representation. I agree that it is counter-intuitive, but it may be right nonetheless. After introducing the HOROR theory I offer some empirical evidence that suggests that we do have conscious experience in the absence of the appropriate first-order states and that disrupting activity in pre-frontal areas (where higher-order representations presumably live) produces a blindsight-like state in normal subjects. Taken together these empirical results show that the HOROR theory is empirically viable despite our intuitions.

If the HOROR theory is right then phenomenal consciousness is nothing but a particular kind of brain activity. This is because phenomenal consciousness is a kind of representation and representations of the right kind turn out to be patterns of synchronized neural activity (possibly in the frontal areas of the brain). What this means is that I will be arguing for what philosophers call a type-type identity theory. The type-type identity theory claims that mental state types are identical to physical state types. After some fancy footwork, which we can avoid here, it can be shown that this amounts to the claim that brains are necessary, and patterns of synchronized neural activity against global chemical background states are sufficient, for consciousness. If the type-type identity theory is right then if there is no brain you do not have mental states or consciousness at all.

Some philosophers might have thought that an identity theory in terms of dynamical states of the brain is inconsistent with any kind higher-order theory, which on their face seem to allow that there may be many ways to have a higher-order representation (in particular ways that are not biological). As in the above case I think it is an empirical question whether we can actually make synthetic or artificial consciousness. It may turn out that something distinctly biological is required for consciousness. If this turns out to be right then we can see our various folk-psychological platitudes about the various kinds of consciousness as a way of picking out or identifying what in the brain we are interested in. On this way of thinking it is the states in the brain that are the representations in question. This kind of view allows powerful responses to various anti-physicalist arguments like Descartes' argument from the conceivability of the distinctness of his mind from his body as well as the more recent arguments based on zombies and the possibility of machine consciousness.

The questions addressed in this chapter are still wide open and there may be many possible routes to physicalism being true. At heart, I am an optimist about the prospects for a complete account of consciousness in physical terms. At the

very least, I hope that the arguments in this chapter can help to support that optimism. I would not go so far as to say that we know that consciousness is a dynamical brain process but I do think that – for all we know – it could be.

2. Some concepts of consciousness

The territory here is at this point well worn so I will provide only a brief exposition of the various concepts of consciousness that I will discuss (for more detailed discussion see Block 1995 & Rosenthal 2005).

Creature consciousness consists in a creature being awake and responding to stimuli. When a creature is unconscious it is not responding to stimuli.

Transitive consciousness consists in our being conscious of things and so consists in sensation, perception, and thinking. A state is transitively conscious when it makes us aware of something in the world. To be aware of something is to be informationally responsive to it.

State consciousness is a property of mental states. Unconscious mental states lack this property and conscious mental states have it. Intuitively a conscious mental state is one that we are, in some way, aware of ourselves as being in.

Access consciousness, as understood by Ned Block (1996), amounts to the idea of a global workspace (Baars 1988; Dehaene & Naccache 2001). That is to say that a state is access conscious when it is broadcasted in such a way so as to be available for the use in reasoning or action.

Phenomenal consciousness is the property of there being something that it is like for one to have a conscious mental state. When a state is phenomenally conscious there is a distinctive way that my experience seems to be. So, when I am phenomenally conscious of, say, a blue patch, there is a particular way that things appear to me. We might say things appear blue but we must also note that the blueness is for me in a particular way. I experience the blue as mine.

This brief survey does not aim to be exhaustive but rather is meant to delineate the topics of the chapter. One concept that I will not discuss in this chapter is the notion of self-consciousness understood as the awareness of oneself as a self.

We can see that there are relationships between these various concepts. For instance, when a creature is conscious it may be in some mental states that are conscious but also be in some mental states that are unconscious. So too when the creature is unconscious it may turn out to be the case that the creature has conscious mental states (perhaps while dreaming). These conscious states may be ones that there is something that it is like for the creature that has them but we must also leave open the possibility that some mental states will not be like anything for the creature that has it.

I will now turn to discussing each of these notions and its relation to the brain.

3. Creature consciousness

Asking the question ‘when is a creature conscious?’ is just asking when is the creature alert and awake? There has to date been a lot of work done on the neurophysiology of sleep and wakefulness. Brainstem areas are implicated in regulating the brain between its waking and sleeping phases by controlling the kinds of neuromodulators that are being released, thereby controlling which neurons are active and inactive. In the awake state the brain is aminergic, meaning that neurons that use histamine, noradrenaline and serotonin are active, and cholinergic, meaning that neurons that use acetylcholine and dopamine are active, in the REM state it is only cholinergic, meaning that noradrenaline, serotonin, and histamine neurons are offline, and in NREM the milieu is chemically intermediate between the two (Hobson 2009: 810).

Hobson and his collaborators have developed a state space model that is very useful. They call it the AIM model (Kahn et al. 1997; Hobson 2009) which allows the state of the brain to be mapped by the amount of activation as measured by global EEG (A), the flow of information as measured by the level of input-output gating (I) and neuromodulatory effects as measured by the excitability of spinal neurons (M). This allows them to distinguish a state space in which we can see waking, sleep, and dreaming each occupying a unique place. For instance when a creature is awake you will find high levels of activation and neuromodulation (indicating that the brain is aminergic) and low values of input-output gating (which is just to say that there is mental activity in response to input and actions being generated). When a creature is asleep we will see a large amount of input-output gating (dreaming of running does not usually lead to moving one’s legs) and low levels of aminergic activity.

Hobson argues that the AIM state space approach provides a natural way to understand states besides waking, non-dreaming sleep, and dreaming. He says,

The ‘state space’ approach also enables the mapping of exceptional mental states such as lucid dreaming and abnormal conditions such as coma and minimally conscious states. Sleep disorders such as narcolepsy and many psychiatric syndromes (such as depression) also find their place in the AIM state space. (Hobson 2009: 810)

We might hypothesize that mood in general can be analyzed in this way. If so then we can see that states of creature consciousness are nothing but global chemical states of the brain. These chemical states are in a constant state of flux throughout the creature’s existence. This is to say that they unfold in time and so can be considered dynamical systems.

Hobson also distinguishes between what he calls primary and secondary consciousness. Primary consciousness on his usage amounts to perception and emotion while secondary consciousness is defined as self-reflective awareness, abstract thinking and metacognition (Hobson 2009: 803). These notions will count as varieties of transitive consciousness according to the present schema since these are all versions of being conscious of something.

4. Transitive consciousness

Transitive consciousness involves being aware of objects (as) in our environment. Thus when we see a red square we are transitively conscious of red and square and perhaps even that the square is red. Transitive consciousness thus involves what we would normally call sensation, perception and thought. This term is unusual but was introduced by David Rosenthal (2005) as a way to capture the fact that this kind of consciousness always involves being conscious of something (the 'of' there is what grammarians would call transitive verb since it takes an object).

The difference between transitive consciousness and creature consciousness highlights a distinction between what we might call states of consciousness and conscious states, which in turn mirrors the distinction between states of the brain and brain states. A state of the brain is a global state which just is the total ratio of aminergic versus cholinergic neuromodulators in the brain at any given moment in time. A conscious state, on the other hand, is a particular state. This is mirrored in the brain by talk about particular brain states. We will be talking about particular representations that can be instantiated against different background states of the brain. Thus we may have the same perception or thought against the very different states of an awake brain versus a dreaming brain.

Wolf Singer is well known for arguing that synchrony in a frequency may be a general strategy that the brain uses to represent various features of objects. In my 2006 paper (Brown 2006) I argued that this general notion could be extended to a hypothesis about the nature of brain states (as opposed to states of the brain). I won't here repeat the experimental data from Singer (1996, 2000), which by this point is well known.

We can extend this idea to offer a viable account of how various mental processes can be understood in terms of synchrony. For instance Gyorgy Buzsaki (1989, 1996) and his lab have argued that we can understand memory in terms of synchrony in various frequencies. Buzsaki assumes that we can understand neural representation in Singerian terms as synchronized neural activity in the gamma range. In the hippocampal formation this gamma activity is modulated by a theta

rhythm that serves to put neurons in 'the appropriate context to receive information'. In essence neurons may 'tune' in to the information that is being broadcast in the theta frequency. At the neural level this happens because the theta rhythm serves to keep the entorhinal neuron's membrane voltage close to but below firing threshold (Buzsaki 1996: p 83). For the full story I would refer the reader to my 2006 paper.

This idea, that some rhythms encode information, or represent things, while others serve as 'carrier' signals on which the information is broadcast and that disparate brain areas can 'tune' in to the information by becoming synchronized in the broadcast frequency, has been recently defended by Edvard Moser and his lab. The gamma frequency comprises a wide swath of frequencies from about 25 hertz all the way to about 150 hertz. The Moser lab has been able to show that neurons can selectively synchronize in either slow or fast frequencies which are themselves phase-locked to different cycles of the theta rhythm (Colgin et al. 2009).

If the foregoing considerations are right, then we can postulate that access consciousness is also nothing but long-range synchronization between different parts of the brain. Neurons in, say CA1 which are firing in synchrony and thereby representing the orientation, say, of something are being broadcast through the hippocampal formation on the theta frequency, meaning that they are phase locked to particular cycles of the theta frequency, then neurons in a later processing stage, CA2, can tune in, or access, that representation by becoming synchronized in the theta frequency, which then it turn disposes those neurons to become synchronized in the gamma range, thereby acquiring the representation.

While in no way conclusive the foregoing empirical theories seem well suited for giving a satisfying physicalistic account of these two fundamental kinds of consciousness in terms of two distinct kinds of dynamic activity of the brain. In the rest of this paper I will argue that the other two notions of consciousness, state consciousness and phenomenal consciousness, can be understood in terms of transitive consciousness. On my view transitive consciousness just is a group of neurons that are firing in sync and this in turn means that the other two kinds of consciousness to be discussed will turn out to be nothing more than this kind of brain activity.

5. State consciousness

At one point in time it was common to assume that all mental states are conscious. Descartes famously argued that the mind was immediately transparent to the person whose mind it was. If I thought, felt, saw, or otherwise experienced

something then I knew that I did. And if I sincerely took myself to be having a thought or experience then I was indeed having that thought or experience. But this is no longer commonplace. We have learned that mental states can occur consciously and that they can occur unconsciously as well. In fact one of the striking discoveries of contemporary cognitive science is just how much of what we do can be done by unconscious processes.

At this point we come to the distinction between first-order and higher-order theories. A first-order theorist, like Fred Dretske (1993), will identify state consciousness with transitive consciousness. A conscious state, they will hold, is one by which I am conscious of something in my environment. One major problem with any first-order view is that there seems to be cases of mental states that are unconscious yet in virtue of which we are aware of something. Classic examples come from priming and masking studies as well as change blindness. In those kinds of cases we have evidence for a mental representation that enables us to perform some task, say completing a word or identifying something quicker, but of which we are completely unaware.

The same kind of problem arises for someone who wants to see state consciousness as merely access consciousness. When we are primed to pick out a red car by being subliminally presented with red, say, the state that represents red is access conscious. It is widely available for control of action (that is why we are primed). Yet it will seem to me as though I saw nothing.

Higher-order theories claim that a mental state's consciousness consists in having a suitable higher-order awareness of being in that state (Armstrong 1968, Rosenthal 2005; Lycan 1996). This amounts to the claim that state consciousness is to be explained in terms of some suitable kind of transitive consciousness. The difference between first-order theorists and higher-order theorists lies in what kind of transitive consciousness is needed. For the higher-order theorist the transitive consciousness must be consciousness of oneself as being in some first-order state. These kinds of theories are divided between higher-order perception and higher-order thought being the right kind of awareness. Hereon we will ignore that distinction.

We cannot settle that debate here but many people find a higher-order theory of state consciousness to be plausible. There is a case to be made that some kind of higher-order theory is part of our common sense thinking about the mind. Intuitively we do not call any mental state of which we are completely unaware a conscious state. If you are completely unaware of believing something what sense is there in calling that state conscious? Granted, the state will be an instance of transitive consciousness, which is to say that it will make me aware of something, but that is not what we mean we talk about state consciousness.

We can have conscious states of consciousness as well as conscious mental states. Thus I can have a mood of which I am not aware myself as being in as well

as a belief of which I am not aware of having. Being aware of these states involves a kind of transitive consciousness and so we can see that state consciousness will turn out to be, on the present view, a kind of synchronized neural activity. In particular it will be the neural activity that is responsible for my being informationally responsive to my own first-order brain states.

6. Phenomenal consciousness

There are many physicalistic theories of phenomenal consciousness but broadly speaking they fall into two categories. There are first-order theorists who see phenomenal consciousness as a particular kind of representation of the world (Tye 2000; Byrne 2001) but there are also first-order theories that see phenomenal consciousness as identical to some brain state even though they go on to deny that the state is representational in any way (Block 1996). The water is muddied here by the view that a state may be representational even if it is not conceptual. I will adopt the inclusive view on which a state can be a representation even if the state involves no concepts. In this sense even Block will think that there are first-order representations of red and that these representations just are what phenomenal consciousness turns out to be.

Thus the question for us is whether we should think of phenomenal consciousness as being identical to first-order representations or higher-order representations. Some philosophers have taken the higher-order theory to be a theory of phenomenal consciousness (Weisberg 2011; Rosenthal 2005). On this view having a conscious pain consists in having a suitable higher-order state that represents the first-order state.

It seems to me that both views are possible and that it is largely a matter of empirical fact which of these turns out to be true. It also seems to me that the balance of evidence is tipped in favor of some kind of higher-order theory. I will develop one such theory that I call the Higher-Order Representation Of a Representation (HOROR) theory of phenomenal consciousness and survey the empirical evidence for it. According to this theory what it is like to consciously experience red is identical to the having of a certain kind of higher-order representation (as usual I assume that the right kind of higher-order representation is one that is seemingly unmediated by inference, etc). This is to say that phenomenal consciousness just is a certain kind of higher-order representation. In particular it is identical to a representation of oneself as having some world-directed (i.e. first-order) representation.

The HOROR theory of phenomenal consciousness has similarities and differences from the traditional forms of both first-order and higher-order theories. It

s similar to a first-order view in that it claims that phenomenal conscious just is a kind of representation. Higher-order theories of phenomenal consciousness claim that the first-order representation of red is not enough for phenomenal consciousness. One needs also to have a higher-order awareness of the first-order state. Thus the traditional higher-order thought and perception views seem to be committed to their being a relationship between the first-order and the higher-order state. This is what has caused some to worry about cases where there is no first-order state (Block 2011). In those cases there would seem to be nothing for the relationship to hold between. This is where the advantages of the HOROR theory come in. According to that view it is not that we must have a relationship between the first-order state and the higher-order awareness. Rather phenomenal consciousness just is the higher-order representation.

The HOROR theory is a theory about phenomenal consciousness. That is, it is a theory about which mental states possess the property of there being something that it is like for the creature to have the states in question. It is not a theory of state consciousness. State consciousness consists in being aware of yourself as being in some state (whether a mental state or a state of mind) or other. It will then be the case that the conscious state is the one of which you are aware of yourself as being in. It is the target of the higher-order representation. But the phenomenally conscious state is not the target of the higher-order representation; it just is the higher-order representation. Thus HOROR theory explicitly denies that any kind of relation is required for phenomenal consciousness. Some versions of the higher-order thought theory do claim this and so this will be a way that the HOROR theory is different from traditional higher-order thought theories.

The higher-order representation is not introspectively conscious – for that it would need to have a third order state targeting it – but it is phenomenally conscious. It is the state in virtue of which there is something that it is like for the subject.

Many people will find the HOROR theory of phenomenal consciousness to be extremely counter-intuitive. Why should we think that phenomenal consciousness just is the having of a representation of a representation? There are roughly two lines of argument that aim to show that this might be the case. The first is the one that Rosenthal has developed (2005) and is based on the role that concepts play in phenomenology and the second is based on recent empirical work from Hakwan Lau's lab (Rahnev et al. 2011).

First we should note the connection between the HOROR theory and phenomenal consciousness. When a state is phenomenally conscious there is something that it is like for me to have that state. It makes sense to think that this entails some kind of awareness must be a part of phenomenal consciousness. If I am experiencing a state as mine, as for me, then there must be something that accounts

for the fact that I do experience it as such. First-order views face the challenge of explaining why some first-order representation would result in there being anything at all that it is like *for* the creature with the representation. If I have a first-order representation of red then I can explain why I am aware of red in the world but it will be a mystery why I take the experience to be mine or for me in the right way. But the HOROR view exploits the resources of higher-order theories to meet this challenge. The specific nature of the higher-order representation involves one representing that one is, oneself, in a particular state. Thus we can explain why a state will be experienced as for the creature. Since one represents oneself as being in a certain state it will seem to you as though you are in that state. Given this we can easily see why having a higher-order representation of a representation would make it the case that there is something that it is like for me. This is a major explanatory advantage of the HOROR theory.

Many will find this picture inviting because there is a straightforward sense in which if one believes that *p* is the case then it will seem to one as though *p*. But many will balk at the step from the claim that it seems as though *p* to the creature to the claim that there is something that it is like for it to seem as though *p*. This is to say that many will think that there is something illicit going on here. There may be a sense in which we can say that it will appear to me as though I am in a certain state by representing myself as being in that state but why should we think that this is the same as the phenomenal sense of appearance? That is, why should we not think that there is a sense in which just having the thought is not enough for the existence of phenomenal consciousness?

Take the case of listening to an orchestra. If one has no concept of what a bass clarinet is one will not consciously experience the sound of the bass clarinet as such, though one's experience of it may be conscious in some other respect (that is to say one will have the relevant first-order states with their qualitative characters and perhaps even higher-order thoughts about them but not as having bass-clarinet* qualities). Once one acquires the concept 'bass clarinet' one's experience is different in a phenomenological way. What it is like for one to hear the orchestra will differ in precisely the sense that it will now sound like there is a bass clarinet in the orchestra to one. The same case can be made for wine tasting. What cases like this give us is data that learning a new concept results in new conscious phenomenology. If concepts can make this kind of difference in our experience then perhaps having them at all can account for the existence of phenomenal consciousness in the first place.

The argument just sketched here may be suggestive but there is a ready response from those who wish to deny the HOROR theory. One may grant that acquiring new concepts is enough to alter one's phenomenology while still denying that concepts can account for the existence of phenomenology. For instance

perhaps the application of concepts actually alters the first-order representations? If one thought that phenomenal consciousness was a property of first-order states, and one thought those states could be changed by conceptualizing them then one could admit that the acquisition of new concepts altered one's phenomenology while denying that phenomenology consists in the application of concepts. This objection may not be decisive but rather than pursue that I will turn to the second, more recent, empirical argument, which is designed to avoid this confound.

This second line of argument is based on the idea that there are empirical cases that seem to suggest that we can in fact have phenomenally conscious experiences in the absence of first-order representations. If we can show that there is genuine phenomenal consciousness without the appropriate first-order activity then we would have strong evidence for some kind of higher-order theory. In other work Hakwan Lau and I (Lau & Brown forthcoming) have discussed three kinds of empirical cases, each serving a slightly different purpose. Here I will focus just on the experimental results from Lau's lab (Rahnev et al. 2009; Rahnev et al. 2011), and I will do this for two reasons. First, I think the other cases we discuss are fairly well understood but the Rahnev et al. results are new and have yet to be fully appreciated. Second, the purpose of each case is slightly different. For instance, we discuss the case of Rare Charles Bonnet syndrome in which subjects with extensive damage to v1 nonetheless report vivid visual hallucinations. This suggests that a particular view about the nature of first-order representations is false (i.e. the view that first-order representations consist in feedback to V1). However it does nothing to show that the first-order representations may not be in other more intermediate brain areas (Prinz 2005). As I will try to show in what follows, the Rahnev experiments are not susceptible to this response.

In this series of experiments subjects were presented with grating patterns or noise in both attended and unattended locations and asked simply whether they saw a target (as opposed to noise) at the probed location. The luminance contrast was adjusted so that performance on the task was matched as between the attended and unattended locations. Using signal detection theory the authors calculated the d' scores, which in effect are simply a measure of how successful subjects are at performing the task. When the d' scores were matched between the two conditions subjects displayed a more liberal detection bias in the unattended location, even though the amount of information, and so presumably first-order representations, was no different (as measured by d'). In a separate condition subjects were asked to discriminate the orientation of the grating (left versus right tilt) and also to judge how visible the stimulus was. Subjects reported higher-visibility ratings for the gratings that were presented in the unattended locations even though they were just as good at discriminating the tilt, as shown by d' .

The above studies use mathematical methods to estimate the amount of information being processed from behavioral data (specifically hits and false alarms). In order to more directly test these issue the authors performed a follow up study using functional magnetic resonance imaging (fMRI). In this follow up study they tracked brain activity in areas that have been implicated in spatial attention. The results showed that when activity in these areas was low, which is thought to correspond to a low state of attention, subjects were more confident that they had seen something in an unattended location. Yet, just as before, their ability to do the task was no better (as measured by d'). In addition to this the authors point out that the average intensity of activity in early visual areas was not higher in either of the two conditions (attended versus unattended). This is what we would expect given that the d' scores are matched. Subjects are performing at the same level and so we would expect to find comparable activity in areas thought to be responsible for first-order representations.

This series of experiments strongly suggests that there can be differences in conscious phenomenology without differences in first-order representations. This is because subjects are telling us that they see something or that it is more visible and yet they are doing no better (or worse) in these cases. If we take their reports at face value then we should allow that gratings in unattended locations are indeed more visible than ones that subjects are attending to. But yet there is no difference in the amount of information being represented by the system (as given by d' and reinforced via the fMRI data from the follow up study). As noted above this suggests that the response from Prinz doesn't affect the Rahnev studies. There appears to be the same amount of information represented in both cases. If there were additional representations in intermediates brain areas in one case but not the other we should not expect to find d' scores that are matched.

It might be objected that this does not actually show that there is a difference in conscious experience since it may be the case that this merely reflects some kind of cognitive bias. However this seems unlikely. In the Rahnev (2011) experiments multiple controls were performed ranging from paying subjects for better performances to giving trial by trial feedback on performance to adjusting the stimulus length/contrast to tracking eye movements to assure that subjects were performing the task correctly. None of these controls destroyed the effect and this suggests that the effect is actually due to perception. If it were merely a cognitive bias then we would expect that it could be trained away. But it is resistant to these kinds of tactics.

The above considerations suggest that there can be conscious experience that does not depend on first-order representations. The next step in the argument aims to show that we have evidence that conscious experience depends on areas of the brain thought to be engaged in higher-order representations. In particular there

is some evidence that the dorsal lateral pre-frontal cortex is the brain area where the higher-order representations can be found. In particular Lau and Passingham (2006) showed that when it is disrupted with TMS bursts subjects report seeing nothing even though they have very good d' scores. These results together suggest that phenomenal consciousness may be higher-order activity in the dorsal lateral pre-frontal cortex and given that we know that it is associated with metacognition it is reasonable to conclude that activity in the dorsal lateral pre-frontal cortex may be the neural substrate of higher-order representations of representations.

Of course the view that we end up with is still something very much like the higher-order thought theory as Rosenthal defends it. It claims that phenomenal consciousness is the having of a certain kind of representation, a higher-order representation of a representation, and further speculates that these representations are in the pre-frontal cortex. All we have done is to accept phenomenal consciousness as a distinct kind of consciousness and to find it a place in the higher-order theory. The HOROR view aims to make clear what the higher-order theory is committed to and what it isn't, but it is a variant of higher-order theory. The main and most notable difference is that the HOROR theory is explicit about the relational requirement applying only to state consciousness. That is to say that state consciousness is explained via a relation of awareness whereas phenomenal consciousness is explained via the awareness itself.

If the forgoing argument is plausible then we have a decent case for thinking that phenomenal consciousness is itself a kind of synchronized neural activity in pre-frontal areas. This is because we have reason to think that phenomenal consciousness is a certain kind of representation and we, in turn, have reason to think that the relevant kinds of representations are nothing but synchronized neural activity in specific frequencies.

The foregoing is a *prima facie* decent case that HOROR theory is empirically viable. Before concluding this section I will note that Ned Block (2007, 2008) and Rafi Malach (2011) have contested the higher-order account on empirical grounds. Block has argued for what he calls phenomenological overflow, which is the claim that phenomenal consciousness outstrips our cognitive access to it. On his view we should think of phenomenal consciousness as neural activity in the relevant first-order sensory areas of the brain rather than as activity in the prefrontal areas. Malach, on the other hand, has suggested that recent empirical work suggests that we can have phenomenally conscious experience when frontal areas are relatively deactivated. I will say a brief word about each before concluding this section.

Malach has used recent results (Goldberg et al. 2006) that suggest that frontal areas responsible for introspection and self-consciousness seem to be relatively inactive while people are absorbed in an external stimulus. Yet we have very good reason from our own cases to expect that one's phenomenal consciousness is very

vivid in these cases. This looks to present a serious challenge to the HOROR approach. However the mistake here is to assimilate these frontal areas with the dorsal lateral prefrontal cortex (Lau & Rosenthal 2011b). The HOROR theory makes no appeal to self-consciousness or introspection and so we would not expect activity in prefrontal areas involved with self-directed introspection to be activated when one is having vivid conscious sensory imagery. HOROR theory is a theory about phenomenal consciousness and so is about our ordinary pre-reflective conscious experiences and not about those rare cases when we turn our attention to our own experiences.

Block uses results from Sperling (1960) and, more recently, work from Victor Lamme's lab (Sligte et al. 2008; Sligte et al. 2009) to argue that phenomenal consciousness is to be identified with activity in first-order visual areas. In the Sperling paradigm subjects are briefly shown an array of letters arranged in a 4×4 grid. Subjects report that they see all of the letters but when asked to name all of the letters they are able to only name about 4. Yet if cued before hand as to which row to attend to they can get most or all of the letters in any given row. Block reasons that if we take them at face value we seem to have evidence that they have more phenomenally conscious experience than they are able to report (after all, they say they see all of the letters but can only report a subset. And the subset could be any row). However it is not at all clear that the arguments for overflow are persuasive at all (Brown 2011). I will not here repeat the arguments but I will just say that both views are compatible with the reports of subjects. If one sees something through a foggy window one may not be able to see all of the details of the object but will no doubt feel as though one has seen the entire object. So too if I am flashed a grid of letters and I see most of them to some extent or other I will be confident that I have seen all of the letters. Given this there is no case for any kind of overflow and so no threat to HOROR theory. The upshot, then, is that the empirical evidence is not strong enough to take overflow seriously, especially when we factor in the independent evidence we have in favor of the non-overflow HOROR view given above (for additional empirical support for the higher-order approach see Lau & Rosenthal 2011a).

I conclude, then, that for all we know HOROR theory is true and phenomenal consciousness just is the right kind of higher-order representation.

7. Identity, reduction, and explanation

As I see things what we have arrived at is a kind of Type-Type identity theory. According to this kind of view types of mental states are identical to types of brain states and types of states of the brain are identical to types of states of mind. So,

depression, on this view, just is a certain range of a dynamic chemical state of the brain, seeing a red bar just is a certain kind of synchronized neural firing. This is because it is plausible that phenomenal consciousness just is a certain kind of higher-order representation and that representation is likely to be synchronized neural activity in dorsal-lateral pre-frontal cortex. However the view we have arrived at is not exactly the same as any of the familiar kinds of type-type identity theory.

Within the identity camp there are two broad traditions that roughly correspond to how one thinks about scientific identities. One view, championed by U.T. Place (2004) and J.J.C. Smart (1991), two of the originators of this theory in philosophical circles, is that mind-brain identities are postulated because they offer the most parsimonious ultimate theory. Thus on this view the postulated identities are brute facts that cannot be explained by anything else. We identify water with H_2O because it allows us to offer the most simple and parsimonious explanation of a wide range of chemical and common sense data.

On the other hand we have a tradition that traces back to David Lewis (1966). On this view the identities are entailed by the theories that make them true. So in the case of water we arrive at the identity of water and H_2O by first identifying water in common sense terms. Water is the stuff that falls from the skies, fills our lakes, etc. We then find out that the stuff that fills our lakes and falls from the sky is H_2O and so we conclude that water is H_2O .

In the foregoing discussion I have been assuming some version of the Lewis strategy. Notice that in each case we started with a common sense way of identifying the kind of consciousness we were interested in and then we found out that that thing turned out to be a particular kind of dynamic brain activity. I favor this kind of view because it allows us to explain why water is H_2O and why consciousness is a certain kind of dynamic brain activity. It also allows us to answer all of the common objections to the identity theory.

Consider first the kind of objections based on conceivability. Descartes famously argued that he could conceive of himself as existing without his body and concluded that he was not his body. The early philosophers who were interested in the brain and came up with the identity theory modeled identity statements as contingent, which means that they just happened to be true but did not have to be true. Just as the fact that we can conceive of Barak Obama losing the election does not show that he is not currently the President of the United States. So too, they reasoned, just because we can conceive of the mind without the body doesn't show that the mind isn't the body in actuality.

The well-known problem with this move is that it seems plausible that scientific identity statements are necessary. Consider one famous philosophical thought experiment known as Twin Earth. Twin Earth is a place where there is

a clear odorless substance that the inhabitants even call ‘water’ which turns out not to be H_2O . Its microstructure is something much more complicated (and which philosophers have chosen ‘XYZ’ to indicate). So, on Twin Earth water is XYZ, not H_2O . This looks like a case where we have something that we might describe as “fool’s water” (Kripke 1980). Fool’s water is stuff that looks like water but is not. In other words there is a strong tendency to think that there is no water on Twin Earth. Water is H_2O , and there is no H_2O on Twin Earth. Others think that there is water on Twin Earth, it just so happens that water – for them- is XYZ.

Given this simple way of thinking about things (for more see Chalmers 2008) we can see that conceiving of a world with a ghostly mind is no problem. Just as the XYZ world did not impugn the identity of water and H_2O in actuality so too the ghost worlds do not impugn the physical credentials of consciousness around here. The ghost world is just another way that we might get consciousness in the world but its conceivability shouldn’t bother us here.

What about zombies? Chalmers (2009) appeals the conceivability of physical duplicates of me that lack phenomenal consciousness. According to the present view the zombie world is akin to a world that is physically identical to our world in that it has H_2O but is stipulated to lack water. This is not even conceivable. Given what we know now we can see that we can in fact deduce water facts from H_2O facts and that shows us that there are no possible worlds like the one described. Just given the H_2O facts alone necessitates water facts. If the mind-brain identity theory is true then the same is the case for mind-brain identities. The zombie world is then inconceivable. What are we to say to the charge that the zombie world seems conceivable? Is this an objection? No. The problem is that it is equally conceivable that consciousness be a physical property. I have previously (Brown 2010) called these creatures ‘shombies’. Shombies are creatures that are completely and exhaustively physical but that are conscious in exactly the same way that I am. In fact what we have seen in this chapter is an argument to the effect that we can conceive of shombies.

One other kind of objection comes from thinking about the possibility of minds that are not composed of neurons. Let us discuss the science fiction example of Commander Data from the Star Trek series. Data is portrayed as having a ‘positronic’ brain, which is supposed to be something like a functional isomorph of the human brain. If it is empirically possible to build something like Commander Data then the type-type identity theory is not true. In that case a kind of functionalism would be true about the mind. This is a possibility but it is an empirical question. If it turns out that there are biological properties of neurons that matter and cannot be reproduced artificially then Commander Data cases will

turn out to be like ghost cases. They will be worlds where consciousness is not a brain state but is rather a positronic state.

As of right now we have no reason to believe in the multiple realizability of consciousness. Instead we have good reason to believe that the mechanisms discussed in this paper hold for all brained species on Earth. We may have intuitions about what could have been the case about consciousness (could it have been positronically based rather than brain based?) but we don't have any empirical reason to think so. So, at least as of now, we cannot take intuitions about machine consciousness as a defeater for the type of reductive view I am arguing for. For all those intuitions machine consciousness just may not be actually possible.

8. Conclusion

The brain is a dynamical system that is constantly evolving in time. There are two faces to this evolution in time. One face is chemical and is the story of how the ratio of aminergic and cholinergic neuromodulators evolves in time. The other face is electrical and is the story of how transient assemblies of neurons are formed via synchronous phase-locked firing and transmitted to disparate areas via long-range synchronization in different frequencies. We discovered these dynamic activities by looking for what in the brain performs various mental tasks. We start with sleep characterized in terms of behavior and then discover the nature of it in the state of the brain. So too we ask how does the brain represent? And we find out that it does so by instantiating a certain pattern of neural activity. It is then natural to deduce that the nature of sensing or sleeping just is the activity in the brain.

This is the way normal scientific identities are established and defended. But it can only be done when one has a theory of the phenomenon that is not couched in neuronal terms. We identify some mental activity as a representation of orientation, say, because we were looking for a representation of orientation. That is, we started with some idea about how we pick those kinds of things out. It is, we think, whatever state we find which reliably tracks this orientation as opposed to some other. It is only because we understand those things in non neural terms that we are able to look at the brain to see what it is that does that. In this way we can see that we cannot have a truly neural theory of consciousness. We must always have some higher-level theory of the thing in question. Neuroscience then is in a position to tell us what in the brain those things are.

Once we recognize that this is the way that scientific identities are discovered we can see that there is no threat to the identity theory from any of the major

objections to it. Neither is there any kind of tension between higher-order theories and biological theories of consciousness.¹

References

- Armstrong, D.M. (1968). *A Materialist Theory of Mind*, London: Routledge.
- Baars, B. (1988). *A cognitive Theory of Consciousness*, NY: Cambridge University Press.
- Block, N. (1995). On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18, 227–47.
- Block, N. (1996). Mental paint and mental latex. In E. Villanueva (Ed.), *Perception*. Atascadero, CA: Ridgeview.
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30, 481–548.
- Block, N. (2008). Consciousness and cognitive access. *Proceedings of the Aristotelian Society*, Vol. cviii, Part 3.
- Block, N. (2011) The higher-order approach to consciousness is defunct *Analysis*, 17(3), 419–431.
- Brown, R. (2006). What is a Brain State?. *Philosophical Psychology*, 19(6), 729–742.
- Brown, R. (2010). Deprioritizing the a priori arguments against physicalism. *Journal of Consciousness Studies*, 17(3–4), 47–69.
- Brown, R. (2011). The Myth of phenomenological overflow *Consciousness and Cognition*. doi:10.1016/j.concog.2011.06.005.
- Buzsaki, G. (1989). Two-stage model of memory trace formation: A role for noisy brain states. *Neuroscience*, 31(3), 551–570.
- Buzsaki, G. (1996). The hippocamo-neocortical dialogue. *Cerebral Cortex*, 6(2).
- Byrne, A. (2001). Intentionalism Defended. In *Philosophical Review*, I: 199–240.
- Chalmers, D. (2008). Two-Dimensional Semantics. In E. Lepore & B. Smith (Eds.), *Oxford Handbook of the Philosophy of Language*. Oxford: Oxford University Press.
- Chalmers, D.J. (2009). The Two Dimensional Argument Against Materialism. In B.P. McLaughlin & A. Beckermann (Eds.), *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press.

1. Some of the ideas in this paper were presented at Columbia University's psychology department as part of their "Cognitive Lunch" Speaker Series under the title "Consciousness and the Tribunal of Experience" March 22 2010. Other ideas in the paper were presented at the Southern Society for Philosophy and Psychology in 2011 as "The Higher-Order Approach to Consciousness: The HOT Ticket or in HOT Water?" and in 2012 as "Phenomenal Consciousness Ain't in the (Back of the) Head". I am grateful to participants for helpful discussion. In particular I have benefitted from discussions with David Rosenthal, Josh Weisberg, Jake Berger, Alex Kiefer, Ned Block, and Pete Mandik. I would also like to especially thank Hakwan Lau for very valuable discussion on many of the ideas in this paper and the data generated by his lab.

- Colgin, L.L., Denninger T., Fyhn, M., Hafting, T., Bonnevie, T., Jensen, O., Moser M. & Moser, E.I. (2009). Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature*, 462(7271), 353–357.
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79, 1–37.
- Dretske, F. (1993). Conscious experience. *Mind*, 102, 263–283.
- Goldberg, I.I., et al. (2006). When the brain loses its self: Prefrontal inactivation during sensorimotor processing. *Neuron*, 50, 329–339.
- Hobson, J.A. (2009). REM sleep and dreaming: Towards a theory of protoconsciousness *Nature Reviews Neuroscience*, 10, 803–813.
- Kahn, D., Pace-Schott, E., et al. (1997). Conscious in waking and dreaming: The role of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience & Biobehavioral Reviews*, 78(1), 13–38.
- Kripke, S.A. (1980). *Naming and Necessity*. Cambridge: Harvard Univ Pr.
- Lau, H. & Brown, R. (Forthcoming). “The Emperors new Phenomenology? The Empirical Case for Conscious Experience without First-Order Representations” in a Festschrift for Ned Block edited by Adam Pautz and Daniel Stoljar (Eds.), MIT Press.
- Lau, H. & Passingham, R. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences* 103(49), 18763–18768.
- Lau, H. & Rosenthal, D. (2011a). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Science* 15(8), 365–373.
- Lau, H. & Rosenthal, D. (2011b). The higher-order view does not require consciously self-directed introspection: Response to Malach. *Trends in Cognitive Sciences*, 15(11), 508–509.
- Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63, 17–25.
- Lycan, W. (1996). *Consciousness*. Cambridge: MA: MIT Press.
- Malach, R. (2011). Conscious perception and the frontal lobes: comment on Lau and Rosenthal. *Trends Cogn. Sci.* 15(11), 507.
- Place, U.T. (2004). Is Consciousness a Brain Process?. In G. Graham & E. Valentine (Eds.), *Identifying the Mind: Selected Papers of U. T. Place*. Oxford University Press.
- Prinz, J.J. (2005). A Neurofunctional Theory of Consciousness. In A. Brook & K. Akins (Eds.), *Cognition and the Brain: The Philosophy and Neuroscience Movement*. Cambridge: Cambridge University Press.
- Rahnev, D., Maniscalco, B., Huang, E. & Lau, H.C. (2009). Inattention Boosts Subjective Visibility: Implications for Inattentional and Change Blindness. *Journal of Vision*, 9(8), 157.
- Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F., Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, 14, 1513–1515.
- Rosenthal, D. (2005). *Consciousness and Mind*. New York: Oxford University Press.
- Singer, W. (1996). Neuronal synchronization: A solution to the binding problem? In Llinas R, Churchland P. *The Mind-Brain Continuum*. The MIT Press.
- Sligte, I.G., Scholte, H.S. & Lamme, V.A.F. (2008). Are there multiple visual short-term memory stores? *Plos One* 3(2), 1–9.
- Sligte, I.G., Scholte, H.S. & Lamme, V.A.F. (2009). V4 Activity predicts the strength of visual short-term memory representations. *Journal of Neuroscience* 29(23), 7432–438.
- Singer, W. (2000). Consciousness from a neurobiological perspective. In: Metzinger T, *Neural Correlates of Consciousness*. The MIT Press.

- Smart, J.J.C. (1991). Sensations and brain processes. In D.M. Rosenthal, *The Nature of Mind*. Oxford University Press.
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74, 1–29.
- Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.
- Weisberg, J. (2011). Abusing the notion of what-it's-like-ness: A response to Block, *Analysis*, 71, 438–443.

An integrative pluralistic approach to phenomenal consciousness

Rick Dale, Deborah P. Tollefsen & Christopher T. Kello

Cognitive and Information Sciences, University of California, Merced /

Department of Philosophy, The University of Memphis / Cognitive

and Information Sciences, University of California, Merced

We review theories and empirical research on underlying mechanisms of selfhood, awareness, and conscious experience. The mechanisms that have been identified for these phenomena are many and multifarious, lying at many levels of space and time, and complexity and abstractness. Proposals have included the global workspace for conscious information, action and its centrality to self awareness, the role for social information and narrative, and more. We argue that phenomenal experience, whatever it “really is,” is probably dependent upon all of these levels simultaneously. We end with two challenges for consciousness research. Both are couched in terms of the dynamics of phenomenal experience. The first is to investigate the sustained dynamics of phenomenal experience; the second is to unveil the way that multi-scale processes in the cognitive system interact to produce that richness of experience. We do not aim to solve the hard problem, but argue that any solution will require this plural characteristic.

1. Introduction: A plural approach

So-called “qualia” remain a mystery. They are a mystery partly because we are not *really* sure if they exist. Some proclaim that they are the very things about which we should be most certain; others proclaim that they are a fanciful illusion, supported only by questionable “intuition pumps” in the form of thought experiments (Dennett 1988). Even if we grant their existence, they seem to be tucked away into first-person, private experience, inaccessible to third-person science. This first-/third-person chasm discussed by Nagel (1974) is still seen by many as an obstacle to progress (e.g. Dietrich & Hardcastle 2004; McGinn 2000); others have forcefully endorsed *both* first- and third-person perspectives as crucial for making progress in a science of consciousness (e.g. Velmans 1990, 2009). Other disagreements abound.

As a neural approach to consciousness was emerging two decades ago, Crick and Koch (1990) sought to overcome such extensive disagreement through a set of conditions. These conditions were intended to keep their discussion from being “frittered away in fruitless argument.” (p. 264). They may have helped kickstart a neurobiology of consciousness, but they did little to stem broader debate. For example, Crick and Koch recommended that researchers should avoid getting stuck in particular definitions of consciousness. Yet some have argued that a clear definition is exactly what we need (e.g. Velmans 2009). Others still argue that a precise definition, or so-called “semantic ascent,” could impede progress when we remain still so confused about consciousness (e.g. Antony 2001). Some have fractionated the concept, identifying more than one kind of “consciousness” – now associated with a large literature itself (e.g. Block 1995, 2005, 2007). As we’ll see below, significant progress has come in our neurobiological understanding by designing basic tasks, based on rudimentary conceptions for now, and taking an incremental approach. The stipulations of Crick and Koch (1990) are instructive to read, as practically each of them remains a continuing issue, from the definition and function of consciousness, to the problem of “qualia.”¹

Despite the continuing debate, the groundwork laid by Crick and Koch (1990) and other researchers has led to progress in the science of consciousness. Part of our goal in this chapter is to review some of this progress, found in impressive work in neuroscience and cognitive science. This large and still-growing literature has the distinct flavor of scientific progress: nascent consensus, thorough experimental investigation, and well-articulated theoretical and quantitative models. The problem, as we will argue, is that our glance at this literature does not support one particular theory of what remains a significant puzzle: phenomenal experience. This is because phenomenal experience is brought about by *diverse* mechanisms. Our background review, we will argue, recommends a *pluralistic* approach to experience. After all, consciousness is a complex biological function underlain by a variety of mechanisms, at a variety of spatial and temporal scales. We will therefore argue that any solution

1. The following is a paraphrase of some key stipulations of their discussion (on their p. 264): (1) Don’t define consciousness precisely, it may mislead us; (2) it is too premature to speculate on the function of consciousness; (3) other animals probably have consciousness (and ergo: language is not necessary); (4) there may be many other forms of consciousness, but few of these (e.g. lucid dreaming) help with scientific headway on the issue; (5) no neural theory will explain everything about consciousness; (6) let’s leave qualia aside for now, though they may be assailable at some point.

to the “hard problem” (Chalmers 1997) will require a *pluralistic* approach to experience.

Admittedly, the observations we offer are not terribly new. They are not meant to be, because they will derive from our background review of such topics as the Global Workspace Theory (Baars 1988; Dehaene et al. 1998), social modulation of cognitive processing and action (Castiello et al. 2010), and the relationship between action and awareness (Jeannerod 2006). We aim, instead, to emphasize that further progress could be made by assembling a more integrative account of phenomenal experience. Our review suggests that phenomenal experience is underlain, sustained, and modulated by *diverse controlling variables* in and outside the organism. We highlight the pluralistic nature of this explanatory strategy: The multi-scale property of phenomenal experience allows *numerous theories* to have a role in the explanation, depending on one’s meaning of “experience,” and what measurement grain size is chosen in analysis. An argument for such “pluralistic” stances can be found in many domains of cognitive science (e.g. Bechtel 1990; Dale 2008; Dale, Dietrich & Chemero 2009; Dove 2009; Edelman 2008; McCauley & Bechtel 2001; Weiskopf 2009). This leads us to a multi-scale account of experience that, we argue, is best articulated through the language of dynamics and complexity science. This is similar in spirit to, and is influenced by, multiple-constraint accounts pioneered by, for example, Metzinger (2003; Wiese and Metzinger, this volume), Merker (2007, this volume), Shallice (1972), Tulving (1985), and of course many more.

2. Outline of the chapter

In what follows, we first cover three key themes in the science of awareness and consciousness, in order of relative grain size, from basic mechanisms to larger-scale social organization: (1) Global Workspace Theory and related discoveries, (2) the role of action in awareness, and (3) the central role of social experience. All three of these sections identify aspects of our nervous system, and more abstract cognitive characterizations, that are part of the story of phenomenal experience. Motivated by the role of all these levels, we then discuss the multi-scale nature of phenomenal experience, and identify two outstanding issues that may be pursued through dynamics and multi-scale complexity. We end by discussing how a more integrative approach would lead to new avenues of investigation into phenomenal experience. We do not aim to solve the hard problem, but argue that any solution would require such a pluralistic agenda.

3. Brains and workspaces

Recent research has made significant headway into what happens in the brain when a person becomes aware of a stimulus. An oft-cited “emerging consensus” derives from Baars’ (1988, 1997) Global Workspace Theory of consciousness, which now has a well-developed neurophysiological description (e.g. Dehaene et al. 1998; Dehaene & Naccache 2001), along with carefully developed behavioral experiments to test it (e.g. Dehaene et al. 2006; Sergent et al. 2005; Sergent & Dehaene 2004), and even large-scale computational models that implement it (e.g. Baars & Franklin 2009; Franklin & Graesser 1999).

Global Workspace Theory identifies consciousness with a process that resembles working memory. At durations of several seconds in length, information becomes broadly accessible to the cognitive system, propagating widely throughout the system to influence action: “a fleeting memory capacity whose focal contents are widely distributed (‘broadcast’) to many unconscious specialized networks.” (Baars & Franklin 2009, p. 166) This has, for example, become associated with a distributed pattern of activity that occurs across regions of the brain when a stimulus is detected – the information rendered by a stimulus propagates throughout the nervous system, engaging it in a kind of “tidal wave” of recurrent activation between low- and high-level systems (among many related proposals, see: Dehaene & Naccache 2001; Del Cul et al. 2007; Tononi & Edelman 1998; Rodriguez et al. 1999; Singer 2001).

A problem that this approach faces, as often noted, is the “hard problem” of consciousness: Why on earth do these processes result in delightful first-person experiences anyway? (Chalmers 1997) For example, if such patterns of neural synchrony, coherence, workspace-neuron connectivity, etc. are *correlated* with consciousness, then a natural question that arises is *why* does consciousness occur through these processes? The query seems intuitive. However, it could cause researchers to further stray. As Dennett (2001) warns, one must remember that the proposal is that this neuronal coherence *is* consciousness. Consciousness is no more than just that (above-threshold) wave of activity that a stimulus engenders when it participates in a conscious experience. This sophisticated (type-type) identity theory aims to avoid a category mistake through what some have called modern cognitive science’s phlogiston (e.g. Churchland 1998): “qualia,” or phenomenal experience, as something else *beyond* these dynamic properties of the nervous system.

But one can still be wary of the overall story so far. This worry does not stem from identifying detection or awareness of some stimulus event with the global neuronal “workspace” (because the data are quite compelling). The problem is that these experiments, and even the theories that motivate them, are based on

very simple conceptions of phenomenal experience. Yoshimi (2004) provides an elegant discussion of this problem, which he calls the “mereological dogma.” Our everyday experience is not constituted by a staccato sequence of stimuli detected flittingly amidst a complex array of background information (see also Spivey 2007). It is not best exemplified by punctuate moments of seeing red patches, or isolated experiences like the taste of beer alone. Everyday experience is suffused in event structure that has a high dimensionality and complex dynamic properties – subtleties that are very difficult to describe in natural language or, at present, any theory. As Yoshimi (2004) argues, “if one begins with *parts* of mental states, one has to outline what might be called a ‘structural neuro-phenomenology,’ which takes account of how conscious states divide into parts, how brain-states divide into parts, and how these two kinds of parts relate to one another.” (See Blackmore 2002 and papers in the same issue for a related discussion, and perhaps disagreement with this point.)

In other words, one could, in a cynical mood, note that most modern cognitive neuroscience of “consciousness” is simply stimulus detection associated with certain kinds of voluntary motor control. The rub is whether you’re happy referring to this as phenomenal consciousness (in a form of identity theory), or believe there’s something left to explain. Many of these researchers would probably accept the basic thrust of this latter concern (see, e.g. Dehaene & Naccache 2001, p. 29). Still, the idea of the global workspace provides clues to how it might work. We revisit this below.

4. Action, awareness, and consciousness

Whatever these internal patterns of activity might be, the nervous system is devoted to engaging its external environment. It does not always sit passively and enjoy conscious experiences of red patches. It acts, and the way it does has important links to phenomenology. Our sense of volitional action relates to our perception of selves, and the distinction between self and other (see Jeannerod 2006 for review). Willful action also brings about changes in experience, such as our sense of relative time between successive events (Haggard et al. 2002). The sense of having caused something in the environment is itself based on a variety of variables that, when satisfied, can experimentally induce illusions of will (Wegner 2003). O’Regan and Noë (2002) and Noë (2005) have laid out convincing arguments and review showing that action is fundamentally related to perceptual processes and experience.

An interesting recent account that further seeks to explain why action is central to consciousness is found in Morsella (2005). In this “supramodular theory,”

phenomenal states emerge in cross-modal and integrative contexts that converge to control body plans. When multimodal constraints converge to bring about executed actions, the accompanying internal states have phenomenal properties that derive from binding these diverse information sources. This is hypothesized to produce a quality to experience that we typically call “consciousness,” “awareness,” and so on.

For example, recent evidence from the first author’s laboratory has shown that explicit predictive actions tend to accompany awareness of learning (Dale et al. *in press*). Predictive actions – reaching into the world in anticipation of where it is going – may reflect a convergence of information from prior perception and action experiences that, in Morsella’s (2005) terms, interfere with ongoing body plans. Phenomenal states reflect the cognitive system’s integration or binding of these experiences for overall skeletomotor control. In this research on learning, Dale et al. (*in press*) found that overt anticipatory behavior significantly correlated with a sense of awareness of that learning. In this sense, the strategy of explicit, stimulus-specific prediction must draw the motor system away from other possible actions and guide it towards a single coordinated activity for a time. Action binds the mind into phenomenal states.

Prediction need not always be conscious, however. There is indeed evidence for unconscious prediction and learning (e.g. Turk-Browne et al. 2010). And overt decision-making and valuation may proceed just fine using a surprisingly flexible “unconscious” part of our cognitive system (Bargh & Morsella 2008). Even in perception, there have been proposals for rapid assemblage of “predictive associations” underlying vision (e.g. Bar 2009). Nevertheless, in a variety of studies, aspects of phenomenal experience have been systematically related to motoric activity, as Morsella (2005) would hypothesize. There is a long literature on distinguishing between “explicit” and “implicit” processes in cognitive psychology that we do not have space to consider here, but it may shed further light on some properties of phenomenal experience, action, and learning (e.g. Cleeremans et al. 1998; Cleeremans & Jiménez 2002; Haggard et al. 2002; Hurley 2002; Reber 1992; Sarrazin, Cleeremans & Haggard 2008; Jordan 1998; Morsella et al. 2008). It is important to note that action, as a specific variable underlying aspects of our experience, may functionally relate to the workspace account: Voluntary action may be related to the process of “broadcasting,” as it could serve the function of coordinating the cognitive system to interact with its world (e.g. Franklin 2000).

5. Sensitive dependence on social conditions

But we do not always act alone. A frequent part of our day-to-day experience has a social dimension. Our thought processes seem to be highly sensitive to these

social variables, such as who is around us when we speak loudly about something, or who is with us during a movie (e.g. Crosby et al. 2008). In the past several years, experimental research in cognitive science has shown a pervasive influence of the social on our thought processes. For example, in the domain of language, when two people converse, they become loosely coupled across many behaviors, from eye movements to gestures to speech patterns and all the way to aspects of discourse, like sarcasm (see Shockley et al. 2009 for a review). This coupling occurs across a variety of spatial and temporal scales, from basic perception all the way to higher cognitive functions. It may be that these interpersonal processes weave together during day-to-day activities, and fundamentally shape human experience.

A popular example of a low-level bridge between persons is the hypothesized mirror neuron system (Gallese et al. 2004; Rizzolatti & Craighero 2004), and research on the network involved in empathy has found similar kinds of patterns (Decety & Jackson 2004) – namely, that humans have emotional circuitry that mediates both the experience of emotion in oneself, and comprehension of emotions in others. There is of course debate about whether processes of behavioral and emotional comprehension are “innate” (Heyes et al. 2005), but one thing is for certain: Human cognition rapidly integrates information to judge states of others.

As an example of this rapidity of processing, inferences based on the beliefs and knowledge of others can sometimes be faster than other types of judgment, such as inferences from signs and symbols (Cohen & German 2010). Though there is debate about whether this skill is innate and automatic (Apperly et al. 2006), there seem to be many robust contexts in which humans are inclined (through learning or otherwise) to carry out fairly complicated judgments of others’ cognitive states. Even basic orienting responses can be influenced by emotion-relevant facial stimuli, especially when they involve potential threat (Friesen & Kingstone 1998; Frischen et al. 2007; Tipples 2006).

So from the perception of action and emotion, to the judgment of “mental states” in other agents, humans are capable of rapidly assessing the social relevance of stimuli at various levels of complexity (Meltzoff & Decety 2003). Cognitive processes involve rapid integration of social information to make, for example, social judgments (Freeman & Ambady 2010), and this process of social judgment and interpretation may work very quickly and robustly with even a minimum of information (Ambady & Rosenthal 1993).

From this necessarily selective review of a vast literature, it may be said that humans are sensitively dependent on social conditions. There is now a growing movement in the cognitive, social, and neurosciences investigating how deeply social variables penetrate mental processes (e.g. see for review: Amodio & Frith 2006; Adams et al. 2010; Balci et al. 2010; Castiello et al. 2010; Frith & Frith 1999; Galantucci & Sebanz 2009; Sebanz et al. 2006; Tomasello 2009).

Given this pervasive influence of “the others,” it may therefore be unsurprising that many have proposed that social information and reasoning, language, and discourse figure into our sense of selves and consciousness. If our cognitive system is actively modeling ourselves in the context of others, then the domains of narrative and discourse may indeed be central to the way we experience everyday life, and conceptualize ourselves in it. Work by Gallagher and others argues that our narrative process helps create the very experiences we have when we navigate the world (Gallagher 2000; see also Harré 2000; Humphrey 1982; and discussion in Dennett 1992).

Narratives help frame our expectations and reactions to events, which partly provide, quite possibly, the richness of human experience itself. They may help us to overcome the problem referred to as the “mereological dogma” described above (Yoshimi 2004) by hinting at answers to “What is the nature of this sense of a continuous self? Is it carried by a succession of momentary minimal selves that are tied together by real connections?” (Gallagher 2000, p. 18). Indeed, Dennett refers to our sense of selves as a kind of “center of narrative gravity” (Dennett 1992), and highlights the potential convenient *fictions* that that we construct for ourselves in these narrative structures (because, after all, centers of gravity themselves are fictions, of an important sort).

In contrast to Dennett’s narrative fictions, Gallagher (2000) reviews work that points to the importance of the narrative structure as “decentered, distributed, and multiplex” (p. 20), granting us insights into mysterious moments of the self like “conflict, moral indecision and self-deception.” (p. 20). Drawing from phenomenology (see also Yoshimi, this volume), Gallagher argues that this decentered conception of self – a self which takes part in a narrative that is heterogeneous, with diverse clusters of narrative structure and activity, containing a variety of important players – is more consistent with empirical evidence and everyday experience than Dennett’s fictional account.

In a recent example that is likely to have an impact on this discussion, Graziano and Kastner (2011) develop a framework within which human consciousness is fundamentally related to social information. The core hypothesis put forth by the authors is that awareness “is a product of social perception” (p. 99), and is supported by a wide range of behavioral and neuroscientific studies. A central part of their explanation is the role of *simulation* in understanding other minds (e.g. more recent reviews: Gallese et al. 2004; Vogeley et al. 2001; and earlier discussion: Goldman 1992). Perhaps most interesting in this paper is the development of explicit, testable predictions about this hypothesis. Graziano and Kastner (2011) offer three such predictions. First, damage to areas associated with social processing (i.e. constructing perceptual models of others) should cause problems with conscious awareness. Second, these same areas

should correlate with activity in brain structures associated with attentional control. And third, the same areas involved in building models of others should also be involved in sustaining models of the self. The authors review current evidence in favor of these predictions, and offer insight on further investigation that would support them.

The idea that consciousness is socially constructed has a simple interpretation that can be easily discarded as false (see discussion in Block 1996; Kurthen et al. 1999), since sociality can be seen as something simply built “on top” of more fundamental processes that engender our phenomenology. But basic processes that distinguish self from other seem foundational even to simple forms of consciousness and selfhood (e.g. Damasio 2010, for recent discussion). And if it is true that any kind of consciousness is somehow emergent from the collection of growing experiences and memories that a person constructs in his or her lifetime, then the social should be central to this, too. Indeed, there are well-developed conceptions of autobiographical memory as driven, centrally, by (the often social) events of our day-to-day lives (Nelson & Fivush 2004). If this is true, then human conscious experience has *as a foundation* social experience and its associated narrative structures.

6. Plural processes underlie phenomenal consciousness

We reviewed a series of proposals for consciousness that range from relatively simple experiences propagated in the global workspace, to processes that lie on longer timescales such as social information and experience. From this review alone, it appears that an account of *human* phenomenal experience is not going to be a simple affair of identifying some key characteristic – some strict sufficiency criterion. The Global Workspace Theory provides some understanding of how cascading interactivity among systems produces experiences of individuated stimulus events; but the extended conscious experience of self in the world may be underlain by narrative structures that depend upon a social timescale. Put simply, all of these proposals help us understand how we experience objects in our world, maintain our sense of agency, and frame it in some broader social and cultural context that provides further hues to our experience.

As Crick and Koch (1990) argued, it should be fairly clear that no theory alone can account for all aspects of phenomenal consciousness. In a discussion of explanatory proposals for consciousness, Dietrich and Hardcastle (2004) provide an impressive list of proposed foundations of conscious experience, and state that all “the items on the list have two properties: they either are necessary (at best), but not sufficient, for consciousness, or are as puzzling as

consciousness itself.” (p. 8) The possibilities reviewed here, while now based on growing empirical evidence and sophisticated theoretical apparatus, may still face these problems too.

Consider, for example, the Global Workspace Theory. In Dehaene et al.’s (2006) elegant threshold account, a considerable amount of local processing of a stimulus can take place “sub-threshold,” and only those stimuli that coordinate a fairly distributed cluster of brain areas are “brought to awareness.” Are these conditions *sufficient* for consciousness? If we granted sufficiency, it would have to be limited in the range of experiences it accounts for. Many acknowledge these limitations, sometimes remarking that any theory must, at present, be provisional.

So the problem is not that any theorist urges their account to be all-encompassing, or to be the necessary and sufficient conditions for consciousness (perhaps some do: Rosenthal 1990; Lau & Rosenthal 2011). We might argue, from the vantage point of admitted dilettantes of this literature, that proposing strict *necessary and sufficient* conditions for complex biological functions is not an enterprise that is likely to be successful. Complex biological functions are underlain by an assemblage of processes, lying at multiple scales, and subject to diverse range of scientific vantage points (see, e.g. Mitchell 2003). An explanation for the complex biological function of consciousness can buck the trend of seeking pure “demarcation” conditions for *phenomenal* consciousness, and instead simply specify how these various processes work together. This presents two challenges that we face to achieve a fuller understanding of phenomenal experience. Both can be framed as an investigation into the *time-extended* dynamics of that experience. The first is articulating the longer-timescale *sustained dynamics* of phenomenal experience. The second is to articulate its many interacting parts.

7. Challenge 1: The sustained dynamics of phenomenal experience

Most of the experimental evidence we have reviewed above is based on brief conscious events in the laboratory. These are, of course, the most easily controlled and induced experimentally. They are punctate – hearing a tone, noticing a word, seeing one interpretation of an ambiguous figure, and so on. One needs such delimiting moments to identify when conscious experiences are occurring so that their effects and neural correlates can be found.

But phenomenal experience does not seem like the staccato sequence of punctate experiences of the laboratory. It has a kind of structure that extends over broader scales of space and time (Yoshimi 2004). How does our cognitive system

sustain this continuity, and produce more complex phenomenal experience?² The science of consciousness will eventually, one would suppose, go beyond these individuated moments, and into the sustained dynamics of day-to-day phenomenal experience. There are two ways this might occur. One could be through studying the “intrinsic dynamics” of the neural system – the continuity of experience arises from some dynamical process that can be identified in the brain. A second is that the dynamics of the brain must be investigated as coupled to continuous flow of energy present in the environment – sustained consciousness is an organism-environment affair.

Taking the second position, Noë (2011) argues that conscious experience must be explained by reference to the way the brain is coupled to the external world. He specifically attacks the notion that our conscious experience can be, fundamentally, identified with processes of the brain. Soft versions of his proposal would be straightforwardly endorsed by many researchers: Ultimately, how a system is coupled to its environment is important to understanding how conscious experience and related phenomena work. However, a much stronger version of this argument is presented in his book, and readers might come away with the impression that recent cognitive neuroscience is so misguided that little insight can be gleaned from more mainstream analyses (such as those described in previous sections of this chapter).

Perhaps among the more interesting issues considered by Noë is that of dreaming. If it is true that conscious experience can be had in dreams, then it seems that the external world is not *purely* necessary in stretches of conscious experience. In other words, the brain is capable of engaging in “closed” dynamics that give way to conscious experience, fully internally in the system itself. But he gives dreams short shrift: “So the appeal to dreams, like the appeal to neuroscientific interventions, leaves us more or less where it starts: with unspecific Cartesian intuitions about the interiority of our experience.” (Noë 2011, p. 180)

The problem with Noë’s discussion is that it assumes opposing positions few researchers would espouse. The appeal to dreams by “brain advocates” need not be for the purpose of showing the *pure and universal* interiority of experience, but rather the fundamental importance of the brain’s own “intrinsic dynamic” (Kello et al. 2007; Van Orden et al. 2003) that permits real conscious experiences (of

2. This question could be answered by some by noting that this phenomenal experience is an illusion (see Blackmore 2002 and the papers in that special issue). But, as sometimes acknowledged by some of these theorists, it still remains a problem to explain the quality of that illusion and why it is there at all. This section could be phrased in these terms too; though the authors of the current chapter have all agreed with each other that such accounts do not succeed in deflating what seems like a “real illusion.”

particular kinds). And indeed, as Noë also observes, when the brain is “recoupled” to the external world, that external world fundamentally changes the landscape of experience through dynamic exchanges with the organism. Yoshimi (2007, 2011, this volume) lays out a mathematical framework for understanding this, and much like Velmans’ (2009) more pluralistic, perspectival approach, sees value in both levels of analysis. In a perspective he calls “active internalism,” Yoshimi argues that intrinsic brain dynamics can be subjected to its own analysis and mathematical modeling. However, in order to understand how it operates in ecological context, we require a conception of how these intrinsic dynamics “fold into” the dynamics of the world itself. He proposes model systems to get at these conceptual questions, focusing on dynamic neural network models.

So Yoshimi would predict that conscious experience can emerge from the intrinsic dynamics of the central nervous system. Indeed, there has been some astonishing evidence gathered in the past few decades about the phenomenal quality of a particular kind of dream referred to as *lucid dreaming*. As observed in Crick and Koch (1990), discussed above, conscious experiences of this kind have not typically figured into scientific discussion of consciousness (but see, e.g. Revonsuo 1995; Metzinger 2003). There is now very strong evidence that lucid dreams are real, and the development of methods by LaBerge et al. (1981) and others has allowed the detection of lucid dreams, and an exploration of their qualities (see Erlacher & Chapin 2010, for a review). In the original work (e.g. LaBerge et al. 1981), subjects identified as lucid dreamers were instructed (before falling asleep) to move their eyes in particular ways during a dream. By measuring oculomotor activity during identifiably REM sleep, researchers were capable of detecting the onset of the (conscious) dream state.

Recent evidence suggests that brain imaging may reveal activities carried out by lucid dreamers. Dresler et al. (2011) instructed lucid-dreaming participants to carry out particular actions during the dream state. They had 6 lucid dreamers first carry out a pre-instructed left-right-left eye-movement so that the onset of the lucid dream could be identified. The participants were then instructed to clench one of their fists 10 times. After this first clenching, they were instructed to do the left-right-left eye movement again, then switch hands and clench once more. Researchers used the eye-movement signal to demarcate regions of brain-imaging data (with fMRI and NIRS) and found the expected lateralized pattern in sensorimotor cortex: They detected the specific activity that was being executed in the dream state.

Other concepts related to conscious experience can also be explored in lucid dreams. In some earlier studies, researchers have investigated the way that lucid dreamers experience time. By having participants count from 1001 to 1010, LaBerge (reviewed in Erlacher & Chapin 2010) found that dream time approximates real

time. Studies since this one have found that there may be modality-specific effects of time perception in lucid dreams. Erlacher and Schredl (2004) found that one particular activity (squats), when performed in dreams, appeared to stretch time relative to the same activity during wakefulness.

These findings show that consciousness during dream states may be similar to that of wakefulness in important ways (e.g. sensorimotor activity and time perception), but that, as Noë would certainly predict, there are important differences. The latter point may seem obvious, given reports by many lucid dreamers that activities such as flying can be used to confirm being in the dream state (see the fundamental role of metacognition in developing and controlling lucid dreaming: Kahan & LaBerge 1994). These studies of cognitive activity in dreams, and the sorts of experiences that occur in them, surely raise many questions, but they serve as evidence, *at least*, that dream consciousness is rich with structure.

Rich conscious experience – a sense of will, and experience of events that can be remembered – all *can* take place in the brain in a relatively self-contained way. One may respond that these phenomenal experiences depend on other already-had interactions with the outside world (e.g. for their content). But this is not the same thing as saying that consciousness *requires* coupling to the external world. The intrinsic dynamics of the human brain are capable of producing distinct qualities of phenomenal experience, with structure that resembles that of world-coupled waking. So, while “distal” explanations of conscious experience must employ active interaction with the external environment, there may be “proximal” accounts of conscious experience that can use the intrinsic dynamics of the human brain as a unit of analysis. As we further discuss in the next section, this relates to the *timescales* that we choose in the explanatory agenda. The slower timescale of brain-world coupling (e.g. social interactions, extended perceptual and event experiences, etc.) serve to constrain the landscape of experiences that occur in the “phenomenological now,” occurring at a faster timescale.

Anderson and colleagues (Anderson 2000; Anderson et al. 2006) have argued that an account of sustained conscious functioning must look to dynamical systems and a property of complex dynamic systems called pink or fractal noise (see also Van Leeuwen & Smit, this volume, for some discussion of this). According to Anderson and colleagues, the operation of the brain gives way to these intrinsic dynamics during conscious processing, which can be identified both in waking states and in REM sleep. Kello (under review; Kello et al. 2010) has developed a model that suggests that sustained operation of a complex nervous system requires this “echo” of noise, as it reflects the balance between two dangerous equilibria: zero activity in the brain (akin to death) and saturation of activity (akin to debilitating seizure) (cf. self-organized criticality: Bak et al. 1987; also its relationship to homeostasis for cognitive function: Parvizi & Damasio 2001). Modes of the

nervous system give way to distinct dynamic patterns, and measurable signatures may reflect the sustained dynamics of conscious experience, from single-neuron firing to system-level activity (for recent discussion see Bieberich 2002; Sevush 2006; Werner 2010).

8. Challenge 2: Multi-scale phenomenal consciousness

So if we suppose that the sustained dynamics of conscious functioning is underlain by a kind of pink-noise-inducing interaction-dominant dynamics, what does it mean? While, as mentioned above, we do not advocate seeking a strict sufficiency demarcation point for conscious vs. non-conscious states, we should still be wary of generic properties that may inadvertently render our explanation relevant to non-biological entities, like traffic jams (Helbing 2001), that violate important explanatory intuitions (e.g. Block 1978). A second challenge we wish to describe is that the dynamics of phenomenal experience involve a richness across levels: They are integrative across space and time, between internal states and the world, and so on.

This is not merely the problem of “binding,” the experimental work of which still focuses on relatively simple phenomenal states like binding features into a single visual percept. Instead, we mean that interaction-dominant dynamics giving way to *sustained* phenomenal experience involves a systematic interplay among different emergent levels of organization (Jordan & Ghin 2006; Van Orden & Holden 2003). It is also not merely a problem of specifying the structure of the nervous system and how dynamic interaction takes place within it (e.g. Buzsáki 2007; Damasio et al. 2000). This is because such exploration of structure does not, by itself, reveal the *functional* properties of that integration and how it gives way to forms of phenomenal experience. Instead, one must combine what can be seen in the interactive dynamics of the brain, in conjunction with behavioral and theoretical explorations of how they sustain our day-to-day experience: What are the *contents*, so to speak, of the various systems-level characterizations, and how are they integrated?³

3. The rapidly proliferating structure-function proposals has been lamented recently, along with efforts to seek integration of these proposals and associated findings (e.g. Morin 2006). See also Saygin et al. (2011) for a promising example of mapping structure to function using advanced brain imaging methods, which may provide future assistance with this general problem.

A range of brain structures and processes have been implicated in levels of awareness and self-consciousness (see Morin 2006 & Seth et al. 2005 for some concise review). The role of the cortex in the global workspace has already been discussed. But a case may even be made for the brainstem, which may play an important role in sustaining a rudimentary form of awareness (see Merker 2007, and this volume, for review). It may do so by *deciding and guiding* for an organism that faces complex probabilistic processing in the cortex, and by integrating that processing with a model of the self as the center of a world with which it is interacting. As Damasio reviews (2010; see also Merker 2007 for review), hydranencephalic children, with almost no cortex but preserved brainstem, indeed appear to show awareness of their environment, despite lacking the rich interconnected neocortex supposedly required by other theories, such as the global workspace.⁴ To theorists such as Merker, *sufficiency* for conscious experience may simply lie in functioning brainstem that can integrate basic awareness of the self with the external environment. Indeed, Sevush (2006) has argued that the dynamics of consciousness are already present in the activity of a single neuron!

But, as Morin responds in discussion of Merker (2007), basic awareness is one component of phenomenology, but does not represent the complete picture. It is likely not sufficient, under other contexts of inquiry, to unveil the nature of phenomenal experience. This could include complex social cognition, which may involve building models of self or other (e.g. Frith & Frith 1999); rapid computation of timing between events of one's actions and those in the world to determine agency (e.g. Wegner 2003); a complex and probabilistic layering of computations to extract perceptual information, and feedback connectivity from conceptual knowledge (e.g. Balceris & Dale 2007); and so on. These levels of organization in the nervous system, associated with particular functions, become integrated under normal circumstances, and bring about the subtle tinges that accompany any complex, extended event.

One way to pursue this strategy is to use more ecological experiential tasks, and collect behavioral and brain data. For example, in widely cited work by Hasson et al. (2004), intersubjective correlation between brains was explored to find what areas are activated during a viewing of *The Good, The Bad, and The Ugly*. They found that, during extended viewing of this film, subjects exhibited a surprising intersubjective consistency in the patterns of activity of the brain. These kinds of

4. The Global Workspace Theory is still consistent with this case, because the dynamic inter-region communication suggested by the global workspace may still be preserved in the case of hydranencephaly. Such a debate is outside of current discussion, but a challenge may be made to *cortex-centered models* of the workspace from these cases.

methods may be employed to explore the *differences* between states of the brain during extended events. Differences in phenomenal experiences may be determined by verbal report (e.g. from what aspects were focused on, accompanied by emotional states, etc.), and these differences may be associated with differing patterns of integration at the systems-level: a kind of “neurocinematics” (Hasson et al. 2008; see also Schier 2009, for a related discussion).

Another means of exploring the sustained and multi-scale dynamics of conscious experience is through computational modeling. In a cleverly entitled paper “Drinking from the firehouse of experience,” Kuipers (2008) describes the development of dynamical systems models for capturing experience. The key computational quality of the proposed model is that the cognitive system must cope with an overwhelming amount of *possible* information from the real-world. The continuous flow of sensory data must be filtered, in some way, to guide behavior. In his model, some of that filtering is done through “trackers” that rely on the external world for such information; other filtering may take place through processes that function like a Kalman filter, which can construct predictive models of internal and external states. The result is, again, a multi-part theory of phenomenal experience, which integrates mind with world, but also proposes a series of internalized computational mechanisms that support this.

We therefore conclude that an agenda to uncover the sustained dynamics of phenomenal experience in all its complexity requires a research agenda that extends the experiencer in time, and can seek an exploration of the interaction among neural events, along with the appropriate behavioral (e.g. eye-movements) and linguistic reports. The methods are diverse. And the discoveries, in our opinion, are likely to reveal pluralistic theoretical relevance: from brainstem, to global workspaces, and social narratives.

9. The hierarchy of human activity: Which scales?

We’ve described the study of consciousness as requiring an integration of multiple scales. Scales in science run the entire gamut of possible measurement, so one may ask which *among* the multiple scales of measurement are the ones relevant to conscious experience? As we consider the multiple scales of consciousness, we should also consider whether a bigger picture is emerging. The parable of the blind men and the elephant fits naturally in a discussion of consciousness (e.g. Sloman & Chrisley 2003), and our concern with scales is no exception. The parts of the elephant are not just different. Importantly, the parts sit at different scales of a hierarchy. As Herb Simon famously wrote, hierarchical structures appear universal to all natural and artificial systems of sufficient complexity (Simon 1973). Humans

are complex in this sense because, for example, neural and other physiological activities at smaller/faster scales hierarchically combine to form behavioral actions at larger/slower scales, which further combine to form social, cultural, and other group phenomena at even larger/slower scales. This description of human activity does not entail reductionism, because phenomena at each scale may require their own ontologies and explanations. This is what is intended by the moniker “interaction-dominant dynamics” (Van Orden et al. 2003), and the recognition that smaller/faster scales are nested within larger/slower scales.

So where does the elephant of phenomenal consciousness sit within the bigger picture of the hierarchy of human activity? If one first considers the temporal extent of phenomenological experiences, then there appears to be a privileged *band* of timescales (in the sense of Newell 1990), roughly on the order of seconds to minutes, at which the elephant sits. People do not experience time in nanoseconds or in years. To be clear, one can hold in mind abstractions of events on any timescale – from the cycling of an atomic clock (over 9 billion cycles per second for the caesium-133 atom; International Systems of Units, 2006) to the deep time of planets and beyond. However, the duration with which humans hold thoughts in mind has a characteristic timescale. The boundaries can be pushed somewhat. For example, some baseball players report consciously perceiving the spin of a major league pitch; some meditators concentrate on a single thought for an hour or more. They still do so under the constraints of the characteristic timescales of phenomenal experience.

So what of our blind men? Each has a piece of the elephant in hand, and the multi-scale, hierarchical perspective gives us a frame in which to relate those pieces. For starters, neural dynamics at the scale of spiking patterns unfold on roughly the same timescale as conscious awareness. Thus one could say that the Global Workspace Theory gives us a functional view of consciousness at the scale of phenomenology (seconds to minutes), and Tononi and Edelman’s (1998) dynamic core hypothesis gives us a neuroscientific view at the same scale.

From here we can go up and down the hierarchy. At a smaller/faster scale, theories of reservoir computing (Maass, Natschläger & Markram 2002) may tell us something about the way that action potentials (spikes) are stitched together to form the dynamic core of phenomenological experience. For the present purposes, reservoir computing refers to the basic idea that recurrent networks with nonlinear dynamics have a generic “fading memory” with a capacity for computation, in the sense that patterns of activity transform as they fade. Studies have shown that the relatively fast timescales of membrane potentials (i.e. their rates of decay, which are on the order of tens of milliseconds) lead to spiking patterns in neural networks that fade on the order of seconds and minutes, in line with phenomenology (see Mauk & Buonomano 2004). Computational studies of homeostatic stability in

spike dynamics further reinforce the *faster* phenomenological timescales, showing how simple neural mechanisms can maintain flexible, ever-changing (metastable) spike patterns, and even relate them to behavior and action (Kello, under review).

Behavior unfolds on multiple timescales, including the timescale of phenomenology. But as we move up the hierarchy, we encounter the slower timescales of learning, long-term memory, culture, and evolution. We leave phenomenology behind, but not consciousness writ large, because slower dynamics shape and constrain faster dynamics. Thus a comprehensive view of consciousness encompasses not only activity on the order of seconds and minutes, but also the larger/slower constraints on that activity. In other words, the larger/slower scales of human activity make spike patterns and behavioral patterns more than just patterns – recall that traffic jams are patterns too, and such patterns often exhibit the same earmarks of complexity (e.g. scaling laws; Kello et al. 2010) as neural and behavioral activity. What traffic jams and many other complex patterns lack, however, is a hierarchical nesting within the dynamics of learning, long-term memory, culture, and evolution.

This challenge posed by a hierarchical, multi-scale perspective is this: What is it about the larger/slower constraints of human activity that endow patterns at the scale of phenomenology with consciousness? In view of complex, self-sustaining and replicating biological organisms, goal-directedness may be an important piece of the puzzle. The basis for this hunch starts with the very definition of pattern formation, which occurs when a physical system runs counter to the second law of thermodynamics, that is, against the universal backdrop of ever-increasing entropy (Deacon 2012). Pattern formation occurs in open thermodynamic systems by virtue of energy moving along a gradient, and gradients are created by placing relatively larger/slower constraints on thermodynamics. For instance, the classic Rayleigh-Bénard preparation uses a stable temperature gradient to induce heat transport in fluid molecules rising from the bottom to the top of a heated pan. Patterns of convection rolls form under the right constraints – that is, a certain gradient coupled with a certain fluid viscosity in a certain container.

The analogy may seem far-fetched, but one conjecture is that the larger/slower constraints of human activity are somehow special, in that they create gradients for neural and behavioral pattern formation that endow them with goal-directedness. And this cascading process of running counter to the second law of thermodynamics predicts a layering of scales, with ever more complex patterns allowing diverse goals, under flexible conditions and constraints, to be carried out (Jordan & Ghin 2006). Human conscious experience may emerge from this; and if it does, it has at its foundation “proliferated” scales of organization, instantiated in the phenomenological band, while being continually constrained by even faster and slower timescales at its fringes. How we accomplish such an integration, and what theories can be integrated, are topics that conclude this chapter.

10. Models of an integrative pluralistic solution

Hierarchical, multi-scale dynamics, expressed in one form or another, have become the centerpiece in many discussions of consciousness (e.g. Chemero 2009; Edelman 2008; Jordan & Ghin 2006; Perlovsky & Kozma 2007; Silberstein & Chemero 2011; Thompson 2007). Though this perspective is gaining recognition, it has not shaped and defined the debate. Instead, even amongst theoretical discussion by many philosophers, consciousness is still frequently discussed in terms of punctate moments, with theories aiming to explain “conscious mental *states*” rather than “conscious mental *processes*.” What we have aimed to do in this chapter is highlight the extraordinary range of processes supposed to be involved: from some core cortical dynamics to the role of higher-level social narratives and related variables. These diverse processes should help us approach two open puzzles about consciousness: the dynamics of sustained conscious experience, and how it is constituted by multiple interacting scales.

At the beginning of this chapter, and after our background review, we argued that a solution to these challenges should utilize an integration of *multiple* perspectives. This strategy is often referred to as “explanatory pluralism.” It is not relativism (“all perspectives are correct”), nor is it a brand of theoretical nihilism (“let’s give up this debate”). Pluralism is inspired by the observation that a complex system, measured in diverse ways, will admit of emergent patterns that “exist” in some (at-minimum observer-centered) way, and that theories capturing these patterns can play a powerful role in our overall understanding of that system (see Dale et al. 2009, for a brief summary). The set of all relevant theories will not necessarily be obviously consistent because, after all, emergent patterns by definition have properties not easily inferable from the properties of the parts that bring about the pattern: More is different (Anderson 1972). The challenge is to have these theories mutually inform one another, rather than seeing them *only* as competitors (McCauley & Bechtel 2001).

So what diverse families of theories are relevant here, and how should they be integrated? We discussed three in this chapter: the Global Workspace, action-centered consciousness, and social constitutivity. But we chose to discuss a particular range of complexity, and used these as examples. Other theories are also relevant. For example, Rosenthal’s well-known higher-order thought (HOT) account may be related to the workspace, because the workspace suggests there is broad informational redundancy in activation across regions of the brain when one becomes conscious of a particular stimulus. HOT would state that these broad patterns of neural communication serve to generate higher-order referential encoding (Lau & Rosenthal 2011). Gathering evidence for this may reveal that HOT is relevant to the Global Workspace, as it may specify one functional role of

cascading activation across the cortex. Comparisons have been drawn between these theories in the past, but it is not obvious to us that they must be seen as mutually exclusive, despite some simplistic comparisons (see Lau and Rosenthal 2011, Table 2).

Another theoretical example is Chalmers' (1997) well-known discussion of protophenomenal panpsychism. In this account, one supposes that matter has both physical *and* phenomenal properties. This perspective is unlikely to convince many people about the phenomenal constitution of the universe's basic physical material. But it may be possible to explore a kind of *neural panpsychism*. Sevush (2006) has recently argued that the neural dynamics that appear to be intrinsic to waking conscious states can be attributed to the firing properties of *single neurons*. In this account, he sees each neuron as having a particular distribution of activity over its dendritic extensions that may constitute a simple form of "consciousness"; as layers of these neurons fire together, their collective dynamics of dendritic activity may be what brings about higher-order, complex phenomenal states. The proposal is interesting, but at the very least it reveals a problem that panpsychism both "suffers from" (Chalmers 2002) yet succeeds in highlighting: There may be levels of complexity cascading across neural activity, each of which may be causally attributable to consciousness, and understanding their *composition* is an open puzzle (what Yoshimi 2004, calls a 'structural neuro-phenomenology').

If consciousness is a complicated biological function underlain by diverse processes at multiple scales of complexity, then we should marshal the tools of dynamics and complexity science to explore it. It therefore seems *very unlikely* that some unitary "theory of consciousness" is going to emerge that renders all other theories irrelevant. Complex biological functions are produced by an assemblage of (only approximately separable) components, each functioning in sophisticated, context-dependent ways. But this may be said for many concepts in cognitive science. Perhaps labeling consciousness the last frontier of cognitive science is getting a bit ahead of ourselves. After all, cognitive scientists are still grappling with fairly vague concepts such as "representation" and "affordances,"⁵ central to many explanations of other cognitive processes, such as language and perception. To move consciousness along, the goal could be to situate it amongst the normal kind of confusion and controversy, characteristic of our young science. One way

5. Anti-representational advocates of "affordances" often state that affordances are somehow better defined than "representations." In the experience of at least one of the authors of this chapter, one can with great facility find as much debate about the nature of affordances as representations; there is, therefore, no *specific* theoretical consensus on either of these ideas from either camp advocating for them.

to accomplish this, in cognitive science, is to go full bore on producing *models* of what we propose to underlie consciousness. The literature on consciousness is replete with verbiage, but relatively short on implemented mechanisms (admittedly, the current chapter is no exception).

Models of how the various emergent capacities of our cognitive system, coupled to the environment, hang together to produce our subjective human experience requires theoretical tools that can carry out such a multi-scale integration. We have described some of these already, including Kuipers (2008) dynamical model. Other hybrid systems may be relevant, such as the large-scale cognitive architecture called LIDA (see Franklin & Patterson, 2006). We might also take a foundational approach, assembling the smaller/faster scales to observe their emergent properties, such as in models of reservoir computing with critical branching (Kello, under review; Kello & Mayberry 2010), K-set models (Kozma et al. 2007), or large-scale models of thalamo-cortical dynamics (Izhikevich & Edelman 2008). Other approaches, such as sequential dynamic systems or “simfrastructures” (Barrett et al. 2006) would allow dynamical modeling that cuts between these bottom-up and top-down strategies. These models permit simulation of large network structures, specified at various scales, producing a generative model of complex system behavior that is still amenable to system-level analysis. The development of adaptive models, with similar goals, is just beginning (e.g. Hernandez et al. 2009; Sanz et al. 2009).

References

- Adams, R.B., Ambady, N. & Nakayama, K. (2010). *The science of social vision* (Vol. 7). Oxford University Press.
- Ambady, N. & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 64, 431–441. doi:10.1037/0022-3514.64.3.431.
- Amodio, D.M. & Frith, C.D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4), 268–277.
- Anderson, C.M. (2000). From molecules to mindfulness: How vertically convergent fractal time fluctuations unify cognition and emotion*. *Consciousness & Emotion*, 1(2), 193–226.
- Anderson, C.M., Lowen, S.B. & Renshaw, P.F. (2006). Emotional task-dependent low-frequency fluctuations and methylphenidate: Wavelet scaling analysis of 1/f-type fluctuations in fMRI of the cerebellar vermis. *Journal of Neuroscience Methods*, 151(1), 52–61.
- Anderson, P. W. (1972). More is different. *Science*, 177(4047), 393–396.
- Antony, M.V. (2001). Is “Consciousness” ambiguous? *Journal of Consciousness Studies*, 8(2), 19–44.
- Apperly, I.A., Riggs, K.J., Simpson, A., Chiavarino, C. & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, 17(10), 841–844. doi:10.1111/j.1467-9280.2006.01791.x.

- Baars, B.J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
- Baars, B.J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA.
- Baars, B.J. & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, 1(1), 23–32.
- Bak, P., Tang, C. & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters*, 59(4), 381–384.
- Balcetis, E.E. & Dale, R. (2007). Conceptual set as a top-down constraint on visual object identification. *Perception*, 36(4), 581–595.
- Balcetis, E.E. & Lassiter, G. (2010). *Social psychology of visual perception*. New York, NY: Psychology Press.
- Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions B*, 364(1521), 1235.
- Bargh, J.A. & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science*, 3(1), 73–79.
- Barrett, C., Eubank, S. & Marathe, M. (2006). Modeling and simulation of large biological, information and socio-technical systems: An interaction based approach. In D. Goldin, S. Smolka & P. Wegner (Eds.), *Interactive Computation: The New Paradigm* (pp. 353–394). Springer Verlag.
- Bechtel, W. (1990). Multiple levels of inquiry in cognitive science. *Psychological Research*, 52(2–3), 271–281.
- Bieberich, E. (2002). Recurrent fractal neural networks: A strategy for the exchange of local and global information processing in the brain. *Biosystems*, 66(3), 145–164.
- Blackmore, S. (2002). There is no stream of consciousness. What is all this? What is all this stuff around me; this stream of experiences that I seem to be having all the time? *Journal of Consciousness Studies*, 9, 5(6), 17–28.
- Block, N. (1978). Troubles with functionalism. In C. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology*. University of Minnesota Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(02), 227–247. doi:10.1017/S0140525X00038188.
- Block, N. (1996). On a confusion about a function of consciousness.
- Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, 9(2), 46–52.
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30(5–6), 481–499. doi:10.1017/S0140525X07002786.
- Buzsáki, G. (2007). The structure of consciousness. *Nature*, 446(7133), 267.
- Castiello, U., Becchio, C., Zoia, S., Nelini, C., Sartori, L., Blason, L., D'Ottavio, G. et al. (2010). Wired to be social: The ontogeny of human interaction. *PLoS ONE*, 5(10), e13199. doi:10.1371/journal.pone.0013199.
- Chalmers, D.J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford University Press, USA.
- Chalmers, D.J. (2002). Consciousness and its place in nature. In S. Stich & T. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 102–142). Blackwell Publishing Ltd. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9780470998762.ch5/summary>.
- Chemero, A. (2009). *Radical embodied cognitive science*. The MIT Press.

- Churchland, P. S. (1998). Brainshy: Nonneural theories of conscious experience. In S. R. Hameroff, A. W. Kaszniak, & A. C. Scott (Eds.), *Toward a science of consciousness II: The second Tucson discussions and debates* (pp. 109–126). Cambridge, MA: MIT Press.
- Cleeremans, A., Destrebecqz, A. & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Sciences*, 2(10), 406–416.
- Cleeremans, A. & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. *Implicit learning and consciousness: An empirical, computational and philosophical consensus in the making*, 1–40.
- Cohen, A.S. & German, T.C. (2010). A reaction time advantage for calculating beliefs over public representations signals domain specificity for “theory of mind.” *Cognition*, 115(3), 417–425.
- Crick, F. & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* (Vol. 2, p. 203).
- Crosby, J.R., Monin, B. & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, 19(3), 226–228.
- Dale, R. (2008). The possibility of a pluralist cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(3), 155–179.
- Dale, R., Dietrich, E. & Chemero, A. (2009). Explanatory pluralism in cognitive science. *Cognitive Science*, 33, 739–742.
- Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L., Parvizi, J. & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience*, 3, 1049–1056.
- Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. Pantheon Books.
- Deacon, T.W. (2012). *Incomplete nature: How mind emerged from matter*. WW Norton & Co Inc.
- Decety, J. & Jackson, P.L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3(2), 71–100. doi:10.1177/1534582304267187.
- Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10(5), 204–211.
- Dehaene, S., Kerszberg, M. & Changeux, J.P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24), 14529.
- Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1), 1–37.
- Dennett, D.C. (1988). Quining qualia. *Consciousness in modern science*.
- Dennett, D.C. (1992). The self as a center of narrative gravity. *Self and consciousness: Multiple perspectives*.
- Dennett, D.C. (2001). Are we explaining consciousness yet? *Cognition*, 79(1–2), 221–237. doi:10.1016/S0010-0277(00)00130-X.
- Dietrich, E. & Hardcastle, V.G. (2004). *Sisyphus’s Boulder: Consciousness and the limits of the knowable* (Vol. 60). John Benjamins.
- Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, 110, 412–431.
- Dresler, M., Koch, S.P., Wehrle, R., Spoormaker, V.I., Holsboer, F., Steiger, A., Sämann, P.G. et al. (2011). Dreamed movement elicits activation in the sensorimotor cortex. *Current Biology*, 21(21), 1833–1837. doi:10.1016/j.cub.2011.09.029.
- Edelman, S. (2008). On the nature of minds, or: truth and consequences. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(3), 181–196.

- Erlacher, D. & Chapin, H. (2010). Lucid dreaming: Neural virtual reality as a mechanism for performance enhancement. *International Journal of Dream Research*, 3(1), 7–10.
- Erlacher, D. & Schredl, M. (2004). Time required for motor activity in lucid dreams. *Perceptual and Motor Skills*, 99(3 Pt 2), 1239–1242.
- Franklin, S. (2000). Deliberation and voluntary action in “conscious” software agents. *Neural Network World*, 10, 505–521.
- Franklin, S. & Patterson Jr, F. (2006). The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent. *pat*, 703, 764–1004.
- Franklin, S., & Graesser, A. (1999). A software agent model of consciousness. *Consciousness and Cognition*, 8(3), 285–301. doi:10.1006/ccog.1999.0391.
- Freeman, J.B. & Ambady, N. (2010). When two become one: Temporally dynamic integration of the face and voice. *Journal of Experimental Social Psychology*, 47, 259–263.
- Friesen, C.K. & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, 5(3), 490–495.
- Frischen, A., Bayliss, A.P. & Tipper, S.P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133, 694–724. doi:10.1037/0033-2909.133.4.694.
- Frith, C.D. & Frith, U. (1999). Interacting minds – A biological basis. *Science*, 286(5445), 1692–1695. doi:10.1126/science.286.5445.1692.
- Galantucci, B. & Sebanz, N. (2009). Joint action: Current perspectives. *Topics in Cognitive Science*, 1(2), 255–259.
- Gallagher. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4(1), 14–21.
- Gallese, V., Keysers, C. & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8(9), 396–403. doi:10.1016/j.tics.2004.07.002.
- Graziano, M.S.A. & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, 2(2), 98–113.
- Haggard, P., Clark, S. & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4), 382–385.
- Harré, R. (2000). Social construction and consciousness. *Investigating phenomenal consciousness*. Amsterdam, Netherlands: John Benjamin Publishing Company, 233–253.
- Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N. & Heeger, D.J. (2008). Neurocinematics: The neuroscience of film. *Projections*, 2(1), 1–26.
- Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634–1640. doi:10.1126/science.1089506.
- Helbing, D. (2001). Traffic and related self-driven many-particle systems. *Reviews of Modern Physics*, 73(4), 1067–1141. doi:10.1103/RevModPhys.73.1067.
- Hernandez, C., Lopez, I. & Sanz, R. (2009). The Operative mind: A functional, computational and modeling approach to machine consciousness. *International Journal of Machine Consciousness*, 1(1), 83–98.
- Heyes, C., Bird, G., Johnson, H. & Haggard, P. (2005). Experience modulates automatic imitation. *Cognitive Brain Research*, 22(2), 233–240.
- Humphrey, N. (1982). Consciousness: A just-so story. *New Scientist*, 95, 473–477.
- Hurley, S.L. (2002). *Consciousness in action*. Harvard University Press.
- Izhikevich, E.M. & Edelman, G.M. (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Sciences*, 105(9), 3593–3598. doi:10.1073/pnas.0712231105.

- Jeannerod, M. (2006). *Motor cognition: What actions tell the self*. Oxford University Press.
- Jordan, J.S. (1998). Recasting Dewey's critique of the reflex-arc concept via a theory of anticipatory consciousness: Implications for theories of perception. *New Ideas in Psychology*, 16(3), 165–187.
- Jordan, J.S. & Ghin, M. (2006). (Proto-) consciousness as a contextually emergent property of self-sustaining systems. *Mind and Matter*, 4(1), 45–68.
- Kahan, T.L. & LaBerge, S. (1994). Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness and Cognition*, 3, 246–264.
- Kello, C.T. (under review). Critical branching neural networks.
- Kello, C.T., Beltz, B.C., Holden, J.G. & Van Orden, G.C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, 136(4), 551.
- Kello, C.T. & Mayberry, M.R. (2010). Critical branching neural computation. Neural Networks (IJCNN), *The 2010 International Joint Conference on* (pp. 1–7).
- Kozma, R., Aghazarian, H., Huntsberger, T., Tunstel, E. & Freeman, W.J. (2007). Computational aspects of cognition and consciousness in intelligent devices. *Computational Intelligence Magazine, IEEE*, 2(3), 53–64.
- Kuipers, B. (2008). Drinking from the firehose of experience. *Artificial Intelligence in Medicine*, 44(2), 155–170. doi:10.1016/j.artmed.2008.07.010.
- Kurthen, M., Grunwald, T. & Elger, C.E. (1999). Consciousness as a social construction. *Behavioral and Brain Sciences*, 22(01), 197–199. doi:null.
- LaBerge, S., Nagel, L.E., Dement, W.C. & Zarcone, V.P. (1981). Lucid dreaming verified by volitional communication during REM sleep. *Perceptual and Motor Skills*, 52(3), 727–732. doi:10.2466/pms.1981.52.3.727.
- Lau, H. & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15, 365–373.
- Maass, W., Natschläger, T. & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11), 2531–2560.
- Mauk, M.D. & Buonomano, D.V. (2004). The neural basis of temporal processing. *Annual Review of Neuroscience*, 27(1), 307–340. doi:10.1146/annurev.neuro.27.070203.144247.
- McCauley, R.N. & Bechtel, W. (2001). Explanatory pluralism and heuristic identity theory. *Theory & Psychology*, 11(6), 736–760.
- McGinn, C. (2000). *The mysterious flame: Conscious minds in a material world*. Basic Books.
- Meltzoff, A.N. & Decety, J. (2003). What imitation tells us about social cognition: A rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 358(1431), 491–500.
- Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30(1), 63–80.
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Bradford Books.
- Mitchell, S.D. (2003). *Biological complexity and integrative pluralism*. Cambridge University Press.
- Morin, A. (2006). Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and Cognition*, 15(2), 358–371. doi:10.1016/j.concog.2005.09.006.
- Morsella, E. (2005). The function of phenomenal states: Supramodular interaction theory. *Psychological Review*, 112(4), 1000–1021.
- Morsella, E., Bargh, J.A. & Gollwitzer, P.M. (2008). *Oxford handbook of human action* (1st ed.). Oxford University Press, USA.

- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- Nelson, K. & Fivush, R. (2004). The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review*, 111(2), 486.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Noë, A. (2005). *Action in perception*. MIT Press.
- Noë, A. (2011). Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness.
- O'Regan, J.K. & Noë, A. (2002). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5), 939–973.
- Parvizi, J. & Damasio, A. (2001). Consciousness and the brainstem. *Cognition*, 79(1–2), 135–160. doi:10.1016/S0010-0277(00)00127-X.
- Perlovsky, L.I. & Kozma, R. (2007). *Neurodynamics of cognition and consciousness*. Springer.
- Reber, A.S. (1992). An evolutionary context for the cognitive unconscious. *Philosophical Psychology*, 5(1), 33–51.
- Revonsuo, A. (1995). Consciousness, dreams and virtual realities. *Philosophical Psychology*, 8(1), 35–58. doi:10.1080/09515089508573144.
- Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192. doi:10.1146/annurev.neuro.27.070203.144230.
- Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B. & Varela, F.J. (1999). Perception's shadow: Long-distance synchronization of human brain activity. *Nature*, 397(6718), 430–433.
- Rosenthal, D.M. (1990). *A theory of consciousness*. Zentrum für interdisziplinäre Forschung.
- Sanz, R., Hernández, C., Gómez, J., Bermejo-Alonso, J., Rodríguez, M., Hernando, A. & Sánchez, G. (2009). Systems, models, and self-awareness: Towards architectural models of consciousness. *International Journal of Machine Consciousness*, 01(02), 255–279. doi:10.1142/S1793843009000359.
- Sarrazin, J. C., Cleeremans, A. & Haggard, P. (2008). How do we know what we are doing? Time, intention and awareness of action. *Consciousness and Cognition*, 17(3), 602–615.
- Saygin, Z.M., Osher, D.E., Koldewyn, K., Reynolds, G., Gabrieli, J.D.E. & Saxe, R.R. (2011). Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience*.
- Schier, E. (2009). Identifying phenomenal consciousness. *Consciousness and Cognition*, 18(1), 216–222. doi:10.1016/j.concog.2008.04.001.
- Sebanz, N., Bekkering, H. & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2), 70–76.
- Sergent, C., Baillet, S. & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, 8(10), 1391–1400.
- Sergent, C. & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, 15(11), 720–728. doi:10.1111/j.0956-7976.2004.00748.x.
- Seth, A.K., Baars, B.J. & Edelman, D.B. (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition*, 14(1), 119–139.
- Sevush, S. (2006). Single-neuron theory of consciousness. *Journal of Theoretical Biology*, 238(3), 704–725.
- Shallice, T. (1972). Dual functions of consciousness. *Psychological Review*, 79(5), 383.
- Shockley, K., Richardson, D. C. & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, 1(2), 305–319.

- Silberstein, M. & Chemero, A. (2011). Complexity and extended phenomenological-cognitive systems. *Topics in Cognitive Science*. doi:10.1111/j.1756-8765.2011.01168.x
- Simon, H.A. (1973). The organization of complex systems. In H. Pattee (Ed.), *Hierarchy theory: The challenge of complex systems* (pp. 3–27). George Brazillier.
- Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1), 123–146.
- Sloman, A. & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, 10, 4(5), 133–172.
- Spivey, M.J. (2007). *The continuity of mind*. Oxford, UK: Oxford University Press.
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
- Tipples, J. (2006). Fear and fearfulness potentiate automatic orienting to eye gaze. *Cognition & Emotion*, 20, 309–320. doi:10.1080/02699930500405550.
- Tomasello, M. (2009). *Why we cooperate*. The MIT Press.
- Tononi, G., & Edelman, G.M. (1998). Consciousness and complexity. *Science*, 282(5395), 1846–1851. doi:10.1126/science.282.5395.1846.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, 26(1), 1–12. doi:10.1037/h0080017.
- Turk-Browne, N.B., Scholl, B.J., Johnson, M.K. & Chun, M.M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, 30(33), 11177–11187.
- Van Orden, G.C., Holden, J.G. & Turvey, M.T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology. General*, 132(3), 331–350. doi:10.1037/0096-3445.132.3.331.
- Velmans, M. (1990). Consciousness, brain and the physical world. *Philosophical Psychology*, 3(1), 77–99.
- Velmans, M. (2009). *Understanding consciousness*. Taylor & Francis.
- Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., Maier, W., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14(1), 170–181.
- Wegner, D.M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences*, 7(2), 65–69.
- Weiskopf, D.A. (2009). The plurality of concepts. *Synthese*, 169(1), 145–173.
- Werner, G. (2010). Fractals in the nervous system: Conceptual implications for theoretical neuroscience. *Frontiers in Physiology*, 1.
- Yoshimi, J. (2011). Phenomenology and connectionism. *Frontiers in Psychology*, 2.
- Yoshimi, J. (2004). Field theories of mind and brain. In L. Embree (Ed.), *Gurwitsch's relevancy for cognitive science* (Vol. 52, pp. 111–129). Dordrecht: Springer Netherlands. Retrieved from <http://www.springerlink.com/content/g568600qt55j1640/>
- Yoshimi, J. (2011a). Active internalism and open dynamical systems. *Philosophical Psychology*, 1–24. doi:10.1080/09515089.2011.569919.
- Yoshimi, J. (2011b). Phenomenology and connectionism. *Frontiers in Psychology*, 2(288). doi:10.3389/fpsyg.2011.00288.
- Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, 6(3), 271–291. doi:10.1007/s11097-007-9052-4.

Index

A

action 5, 43, 47, 51, 152,
155–158, 160, 165, 174, 176,
214, 217, 233–237, 242, 245,
248–249
selection 37
activation 13–14, 249–250
mutual 28
recurrent 234
space, neural 4, 27
state 64, 214
activity, neural 5, 7, 11, 13,
21–22, 25, 28, 48, 57, 82,
88, 100, 132–141, 168, 211,
223, 234, 239, 245
clustering of 8
reverberatory 29
trajectory space 109–117,
194
sustained 31–33
agreement, global 87, 89
ambiguity 21, 26, 28,
37–44, 140
anesthesia 109
assembly, neuronal 21, 27, 29,
82, 200
attractor 124, 126, 137–138,
154–161
dynamics 86–90
limit cycle 155
network 45
point 200
autonomy, of experience 81–83
awareness
vs. consciousness 150
see also self-awareness

B

backward masking 31
Balint syndrome 33–34
best estimate
sensory 56, 67
see also buffer, best estimate

binocular rivalry 31
body-world distinction 37,
54, 69
brain state, *see* activation state;
activity, neural
brainstem 60, 214, 245
buffer
best estimate 37, 46, 51,
55–59, 66–69
Byzantine generals problem 87

C

capacity, representational 109,
111–117, 194
category 152
discrimination 221–222
fuzzy 234
visual 26
causality 187
causal nexus 105
change blindness 57, 159, 217
chaotic itinerancy 88, 137
cluster analysis 8–12
coherence interval 140–141
combinatorial explosion 26
communication, in the
brain 86–87, 138, 141, 245
complex, extended
intralaminar 60–61,
67–70
complexity 16, 124, 127,
132–133, 150, 193, 233
combinatorial 82
of activity trajectories
110–117
computation
digital 96–117
digital vs analog 114
connectivity 100–101, 234
cortical 44–45, 63, 152
local 28–30
consensus, neural 28–32,
82, 87

constraint satisfaction 37, 55,
58, 63–64, 67
interdisciplinary
methodological 185, 207
continuity
strong global 189
temporal 57, 167, 185,
197, 241
coordination, distributed
dynamic 67, 87–90
cortex, cerebral 25, 29–33,
38, 41, 43–47, 54, 58–69,
139, 245
entorhinal 34
frontal 33–34, 223–225
medial-temporal 34
parietal 33
counterfactuals 82–85
creativity 123
Cyclopean 49, 51, 53, 55, 64

D

diagnosis theorems 89–90
dimensionality 60, 235
effective 110–111, 168
distributed system,
asynchronous 87
dream states 24, 66, 186, 188,
213–215, 241–243
lucid 232, 242–243
dwell times 125
dynamical systems 3–4, 16,
85–86, 89–90, 106, 114,
160, 211, 227, 243
analog 56, 108, 117
diagnosing, *see* diagnosis
theorems
electrochemical 99–101
fuzzy 149
double well model 125
open 56
symbolic 161
theory 165–167

- E**
 efferent copy 33
 egocentric reference frame 52, 55, 64
 emotion 160, 168–169, 176, 215, 237, 246
 entropy 159, 248
 environment 2, 33, 51, 86, 88, 121, 152, 156, 161, 181, 235
 event-related potential (ERP) 129
 explanation 82, 115, 211, 224, 233
- F**
 feedforward processing 63
 firing rate, instantaneous 28
 flash suppression 65
 functional magnetic resonance imaging (fMRI) 4–9, 13–14, 135, 222, 242
 functional organization 29, 38, 69, 96–98
 fuzzy
 category 153
 dynamics, *see* dynamical systems, fuzzy
- G**
 global
 unity 186–187
 Global Workspace
 Theory 200, 213, 233–235, 239–240, 245–249
 Globalization by
 Representational Fusion, Principle of 202
 grandmother neurons 26
 Gurwitsch, Alexander 166, 175
- H**
 hallucination 24, 221
 hierarchy, visual 25
 higher-order 16, 60, 167, 250
 representation of a
 representation theory (HOROR) 221–228
 theory (HOT) 217, 249
 hippocampus 34, 45, 215
 holism 44, 86, 185–188, 191
 holon, phenomenal 186
 homology, multiscale 111–112, 116
- horizon, of
 understanding 172–176
 Husserl, Edmund 1–5, 17, 165–183
- I**
 identity 196
 diachronic 188, 205
 theory, type-type 212
 ignition, neuronal 28–35
 illusion
 of will 235
 visual 23
 illusory contours 23
 imaging, brain 31, 65
 see also fMRI
 impossible figures 23
 inattention blindness, *see* change blindness
 independent component analysis (ICA) 7
 infancy 152
 information, integrated 35, 141
 invariance
 organizational, principle of 96, 102, 117
 perceptual 54, 178, 190
 isomorphism 84, 102, 114, 116, 226
- J**
 James, William vii, 53, 81, 154, 167
- K**
 Kant, Immanuel 166, 171, 176, 183, 221
- L**
 laminar periods 135
 learning 151, 156, 236, 248
 lesion 29, 33, 69
 linear algebra 83
- M**
 machine sentience 95
 meditation 159
 mereology 168
 mereotopology 190–194
 meta-representation 192–193, 198, 201, 203–207
 metastability 90, 248
 microstates, brain 133
- memory 2, 24, 54, 57–58, 64, 66, 126, 154, 156, 215, 234, 246
 monism 22
 motor output 32, 46–47, 52, 59, 67, 155, 177, 235–236
 multidimensional scaling (MDS) 7–8
 multi-scale
 dynamics 246
 homology 111–112, 117
 structure 113, 118, 231, 233
 multistability 123
 music
 mind as 87–88
 orchestra, listening to 220
- N**
 narrative 238
 social 231, 246
 neglect 33, 65–66, 68
 neural chauvinism 101
 neural correlates of
 consciousness (NCC) 16, 175, 240
 neural fatigue 126
 neural replacement
 scenario 97–102
 noise 43, 88, 124–127, 136, 221, 243
 fractal 126–127, 135, 243
 1/f 127, 135
- O**
 oscillations
 alpha 124, 131–132, 136–141
 beta 132, 138, 140–141
 delta 66, 132
 gamma 31, 131, 138–139, 215–216
 theta 132, 215–216
- P**
 panpsychism 84, 250
 perceptual sniff 126
 phenomenal “now”
 189, 191–192, 197, 203, 205, 207
 phenomenology,
 constitutive 165
 physicalism 212
 physics 83–84, 96, 102
 Plato 176

prediction 199–203, 236
 probability distribution
 37, 40–43, 45, 54, 59–60,
 62, 66, 174
 protention 2, 12–17
 pulvinar 37, 60–70

Q

qualia 160, 231–232, 234

R

readout 27, 87
 real time 82
 realism, naive 55
 recurrent processing 18,
 28–29, 124, 234
 reduction, explanatory 221,
 224, 247
 relativity theory, special 83
 retention 2–4, 12–17

S

Sartre, Jean Paul 1, 17
 self-awareness 17
 self-embedding, global
 192–193, 198
 self-model, phenomenal 192
 self-observation 86–87, 90
 self-representation 193,
 198, 200
 system 203
 sentience
 machine, *see* machine
 sentience
 signaling, neuronal 28, 86
 silent neurons 35, 85

simulation 102–110, 238
 digital 96–98, 112–117
 neural 54
 sky, dome of 50
 sleep 33, 66, 109–110, 158, 167,
 214, 227, 242
 social information 231
 space, multidimensional 83
 specious present vii, xi, 189,
 203, 207
 speed of light 83
 split brain 30
 stability 124, 127, 138, 159, 247
 see also multistability
 state indicator function 111
 states, experiential 82, 109–112
 stream of consciousness vii,
 155, 167, 169, 192
 structure, surplus 123–124, 141
 style, perceptual 132
 superior colliculus 45–47,
 58–60, 64, 67–70
 support vector machines
 (SVM) 13–18
 synchronization 125, 133–138,
 216, 227
 dynamic 133
 phase 134
 synfire 85

T

Takens theorem, *see* diagnosis
 theorems
 taxonomy, of
 consciousness 213
 template 25–26

time

 continuous vs discrete
 114, 116
 scale 151–152, 206, 239
 timeliness 81–82
 thalamus
 dorsal, association nuclei 60
 pulvinar 37, 60–70
 thermodynamics 84, 248
 threshold, consciousness 31,
 138, 234
 topology 107, 191
 tractability 81–82, 87–90
 trajectory space
 of neural activity, *see*
 activity, trajectory space
 cylinder sets in 152–159
 transmission delays 85
 transcranial magnetic
 stimulation (TMS) 24,
 27, 223
 Twin Earth 225–226

U

uncertainty 37, 40–42, 139
 see also ambiguity
 unconscious 37–38, 42, 48, 58,
 66, 213, 217, 234, 236

W

whole parts 187
 wisdom 156

Z

zombies 212, 226
 zona incerta 61, 67–68, 70