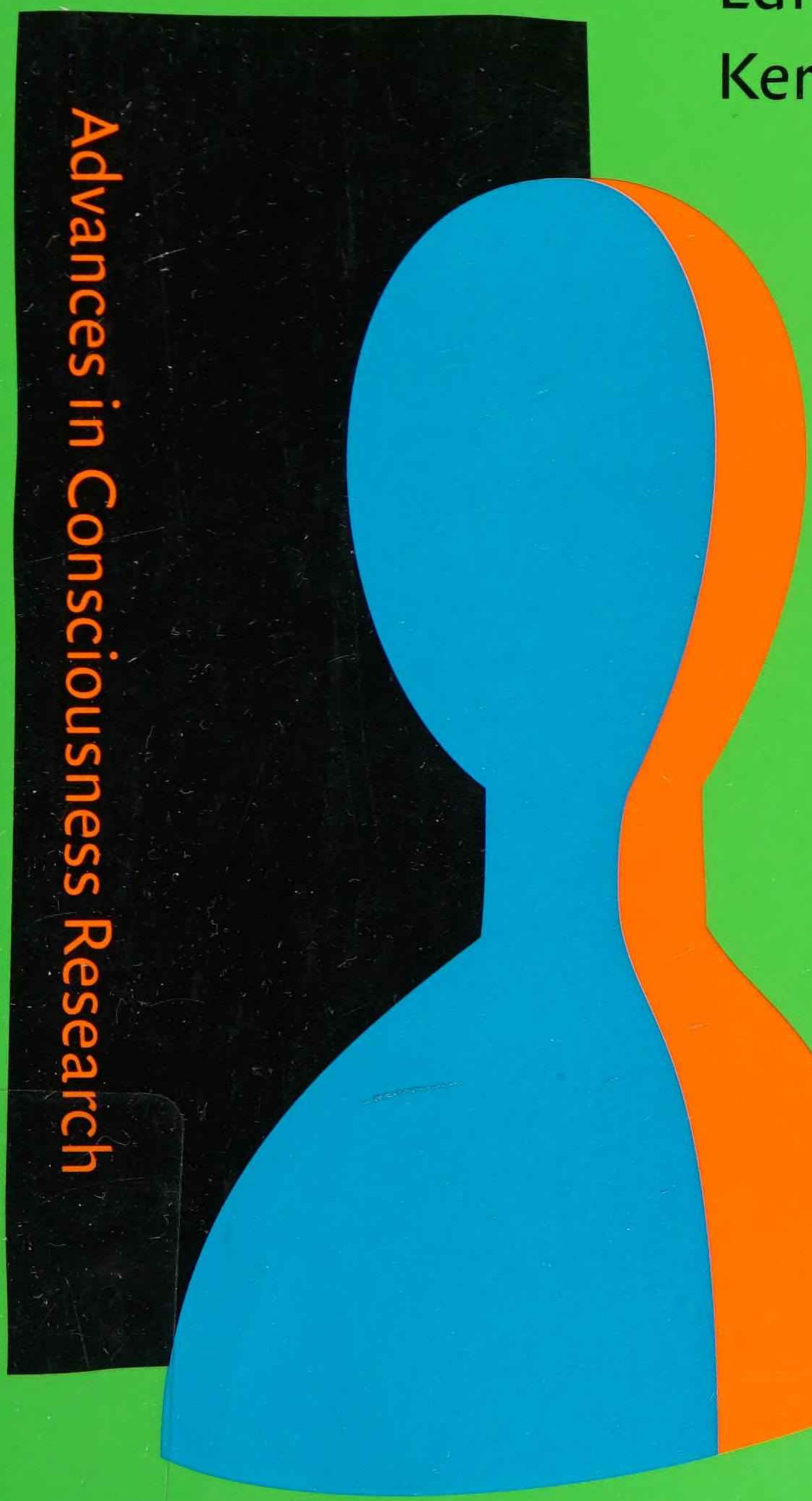
Human Cognition and Social Agent Technology

Edited by
Kerstin Dautenhahn



NO LONGER PROPERTY OF
FALVEY MEMORIAL LIBRARY

FALVEY MEMORIAL LIBRARY
VILLANOVA UNIVERSITY
VILLANOVA, PA 19085

HUMAN COGNITION AND SOCIAL AGENT TECHNOLOGY

ADVANCES IN CONSCIOUSNESS RESEARCH

ADVANCES IN CONSCIOUSNESS RESEARCH provides a forum for scholars from different scientific disciplines and fields of knowledge who study consciousness in its multifaceted aspects. Thus the Series will include (but not be limited to) the various areas of cognitive science, including cognitive psychology, linguistics, brain science and philosophy. 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 and society.
Series B: Research in Progress: Experimental, descriptive and clinical research in consciousness.

EDITOR

Maxim I. Stamenov
(*Bulgarian Academy of Sciences*)

EDITORIAL BOARD

David Chalmers (*University of Arizona*)
Gordon G. Globus (*University of California at Irvine*)
Ray Jackendoff (*Brandeis University*)
Christof Koch (*California Institute of Technology*)
Stephen Kosslyn (*Harvard University*)
Earl Mac Cormac (*Duke University*)
George Mandler (*University of California at San Diego*)
John R. Searle (*University of California at Berkeley*)
Petra Stoerig (*Universität Düsseldorf*)
Francisco Varela (*C.R.E.A., Ecole Polytechnique, Paris*)

Volume 19

Kerstin Dautenhahn (ed.)

Human Cognition and Social Agent Technology

HUMAN COGNITION AND SOCIAL AGENT TECHNOLOGY

Edited by

KERSTIN DAUTENHAHN
University of Reading

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM/PHILADELPHIA

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

Library of Congress Cataloging-in-Publication Data

Human cognition and social agent technology / edited by Kerstin Dautenhahn.
p. cm. -- (Advances in consciousness research, ISSN 1381-589X ; v. 19)
Includes bibliographical references and index.
1. Cognition. 2. Learning. Psychology of. 3. Socialization. 4. Artificial intelligence. 5. Technology--Psychological aspects. I. Dautenhahn, Kerstin. II. Series.
BF311.H766 1999
153--dc21 99-39726
ISBN 90 272 5139 8 (Eur.) / 1 55619 435 8 (US) (Pb; alk. paper) CIP

© 2000 – 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 75577 • 1070 AN Amsterdam • The Netherlands
John Benjamins North America • P.O.Box 27519 • Philadelphia PA 19118-0519 • USA

This book is dedicated to my parents Annelie and Claus-Peter Dautenhahn with gratitude for their love and support over the years.

Kerstin Dautenhahn

Digitized by the Internet Archive
in 2022 with funding from
Kahle/Austin Foundation

https://archive.org/details/humancognitionso0000unse

Table of Contents

About the Editor ix

List of Contributors xi

Introduction xv
Kerstin Dautenhahn

CHAPTER 1 1
Narrative Intelligence
Phoebe Sengers

CHAPTER 2 27
Digital Augmentation of Keepsake Objects: A Place for Interaction of Memory, Story, and Self
Jennifer Williamson Glos

CHAPTER 3 61
Children as Designers of Interactive Storytellers
"Let me tell you a story about myself..."
Marina Umaschi Bers, Justine Cassell

CHAPTER 4 85
Autonomous Synthetic Computer Characters as Personal Representatives
Linda Cook, Tim Bickmore, Sara Bly, Elizabeth Churchill, Scott Prevost, Joseph W. Sullivan

CHAPTER 5 113
"Conscious" and Conceptual Learning in a Socially Situated Agent
Myles Bogner, Uma Ramamurthy, Stan Franklin

CHAPTER 6 137
Emotionally Grounded Social Interaction
Dolores Cañamero, Walter Van de Velde

CHAPTER 7 163
Architectural Requirements for Human-Like Agents Both Natural and Artificial: What sorts of machines can love?
Aaron Sloman

CHAPTER 8 197
Connecting Reflection and Action: A Heterogeneous Multi-Agent Model
Ruth Aylett, David Barnes

CHAPTER 9 225
The Role of Evaluation in Cognition and Social Interaction
Maria Miceli, Cristiano Castelfranchi

CHAPTER 10 263
The Ontogeny of the Social Self. Towards a Formal Computational Theory
Eric Werner

CHAPTER 11 301
Computational Embodiment: Agents as Constructed Complex Systems
Christopher Landauer, Kirstie L. Bellman

CHAPTER 12 323
Are We Having Fun Yet? Using Social Agents in Social Domains
Leonard N. Foner

CHAPTER 13 349
The Emergence of Personality: How to Create Souls from Cells
Steve Grand

CHAPTER 14 377
Machine-Mediated Communication: Agents of Representation
Bill Vorn

CHAPTER 15 395
Agents as Artworks and Agent Design as Artistic Practice
Simon Penny

CHAPTER 16 415
Living with Socially Intelligent Agents: A Cognitive Technology view
Kerstin Dautenhahn, Chrystopher L. Nehaniv

Subject Index
Name Index

About the Editor

Kerstin Dautenhahn is Lecturer in Cybernetics at the University of Reading, United Kingdom. She received a doctoral degree in 1993 for research on human path planning from the Department of Biological Cybernetics, University of Bielefeld, Germany. Her thesis combined psychometrical experiments with computer simulations. From 1990 to 1992 she was employed as a researcher in the Faculty of Mathematics at University of Bielefeld. In 1993 she joined the German National Research Center for Information Technology (GMD), Sankt Augustin, Germany as a post-doctoral researcher in the Artificial Intelligence (AI) division. Since 1993 she has been working on the design and control of mobile robotic systems. In 1996 she became a senior researcher at Vrije Universiteit Brussel (VUB), AI Lab, funded by a European HCM/ TMR research grant. In 1997 she moved to England to her current position in the Department of Cybernetics at the University of Reading, where she teaches "Artificial Life", "Animal Systems", and "Mobile Robotics".

Her main research interests are in the area of socially intelligent agents, comprising the study of different kinds of agents (virtual, robotic and biological). Her practical work focuses on developing designs, scenarios, implementations and applications for agents which can interact and communicate with each other and with humans. Her work and publications address philosophical and cognitive science issues, aspects of system design and control, evaluation of experiments, as well as the process of how to make artificial "life-like" systems believable and useful to humans. Particular topics which Dr. Dautenhahn has done research in are mobile robotics, social learning, and imitation. Her work has combined research into building and designing agents (including robots) with the investigation of cognitive issues, such as the question of how social intelligence can be grounded in dynamic agent-agent interactions, e.g. by studying imitation and the role of narrative in social intelligence. She is very active in organising meetings on topics which are related to her research interests, has served on the programme committees of various workshops and conferences, has been invited plenary speaker at

international scientific meetings, and is a member of the Board of Governors of the Cognitive Technology Society. Additionally she frequently guest edits journal issues on topics relevant to socially intelligent agents research, and she is Associate Editor of the journal *Adaptive Behavior*.

Contributors

Ruth Aylett
Centre for Virtual Environments
Business House
University of Salford
Salford
M5 4WT
United Kingdom
Email: r.s.aylett@salford.ac.uk

Dr. Dave Barnes
University of Wales, Aberystwyth
Department of Computer Science
Penglais
Aberystwyth
Ceredigion
SY23 3BD
Wales
United Kingdom
Email: dpb@aber.ac.uk

Dr. Kirstie L. Bellman, Principal Director
Aerospace Integration Science Center
The Aerospace Corporation, Mail Stop M6/214
P.O.Box 92957
Los Angeles, California 90009-2957
USA
Email: bellman@aero.org

Tim Bickmore
MIT Media Lab
Room E15-320
20 Ames St.,
Cambridge, MA 02139
USA
Email: bickmore@media.mit.edu

Sara Bly
Sara Bly Consulting
24511 NW Moreland Road
Hillsboro, OR 97124,
USA
Email: sara_bly@acm.org

Myles Bogner
Institute for Intelligent Systems
Department of Mathematical Sciences,
Box 72
The University of Memphis
Memphis, TN 38152
USA
Email: bognerm@msci.memphis.edu
URL: http://www.msci.memphis.edu/~bognerm

Dolores Cañamero
IIIA-CSIC
Campus de la U.A.B.
E-08193 Bellaterra, Barcelona
Spain
Email: lola@iiia.csic.es
URL: http://www.iiia.csic.es/~lola

Justine Cassell
MIT Media Laboratory
20 Ames Street
Cambridge, MA
USA
Email: justine@media.mit.edu

Cristiano Castelfranchi
Institute of Psychology of the National Research Council of Italy

Division of Artificial Intelligence, Cognitive Modelling, and Interaction
Viale Marx 15
00137 Rome
Italy
Email: cris@pscs2.irmkant.rm.cnr.it

Linda Cook
FX PAL
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA, 94304
USA
Email: 1_cook@mindspring.com

Elizabeth Churchill
FX PAL
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA, 94304
USA
Email: churchill@pal.xerox.com

Kerstin Dautenhahn
Department of Cybernetics
University of Reading
Whiteknights PO BOX 225
Reading, RG6 6AY
United Kingdom
Email: K.Dautenhahn@cyber.rdg.ac.uk
URL: http://www.cyber.rdg.ac.uk/people/kd/WWW/home.html

Leonard N. Foner
MIT Media Lab
20 Ames St, E15-305
Cambridge, MA 02139
USA
Email: foner@media.mit.edu
URL: http://www.media.mit.edu/~foner/

Stan Franklin
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152, USA
Email: stan.franklin@memphis.edu
URL: www.msci.memphis.edu/~franklin

Jennifer W. Glos
Third Age Media
585 Howard Street
San Francisco, CA 94107
USA
Email: jenglos@alum.mit.edu
URL: http://jenglos.www.media.mit.edu/people/jenglos

Steve Grand, Director
Cyberlife Research Institute
North Down Lane, Shipham
BS25 1SL
UNITED KINGDOM
Email: steve.grand@cyberlife-institute.co.uk

Dr. Christopher Landauer
Aerospace Integration Science Center
The Aerospace Corporation, Mail Stop M6/214
P.O.Box 92957
Los Angeles, California 90009-2957
USA
Email: cal@aero.org

Maria Miceli
Institute of Psychology of the National Research Council of Italy
Division of Artificial Intelligence, Cognitive Modelling, and Interaction
Viale Marx 15
00137 Rome
Italy
Email: maria@pscs2.irmkant.rm.cnr.it

Chrystopher Nehaniv
Interactive Systems Engineering
University of Hertfordshire
College Lane
Hatfield Herts AL10 9AB
United Kingdom
Email: c.l.nehaniv@herts.ac.uk

Simon Penny
Associate Professor of Art and Robotics
CFA 300 Carnegie Mellon University
5000 Forbes Ave Pittsburgh, PA 15213
USA
Email: penny+@andrew.cmu.edu

Scott Prevost
FX PAL
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA, 94304
USA
Email: prevost@pal.xerox.com

Uma Ramamurthy
Dept. of Biostatistics and Epidemiology
St Jude Children's Research Hospital
332 North Lauderdale,
Memphis, TN 38105,
USA
Email: urmmrthy@memphis.edu
URL: http://www.msci.memphis.edu/~ramamurt/

Phoebe Sengers
Media Arts Research Studies
Institut für Medienkommunikation
GMD Forschungszentrum
Informationstechnik
Schloss Birlinghoven
D-53754 Sankt Augustin
GERMANY
Email: Phoebe.Sengers@gmd.de

Aaron Sloman
Professor of AI and Cognitive Science
School of Computer Science,
The University of Birmingham,
Birmingham, B15 2TT
United Kingdom
Email: A.Sloman@cs.bham.ac.uk
URL: http://www.cs.bham.ac.uk/~axs/

Joseph W. Sullivan
FX PAL
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA, 94304
USA
Email: sullivan@pal.xerox.com

Marina Umaschi Bers
MIT Media Laboratory
20 Ames St. E15-320A
Cambridge, MA 02139
USA
Email: marinau@media.mit.edu
URL:http://marinau.www.media.mit.edu/~marinau

Walter Van de Velde
Riverland Next Generation, Starlab
Excelsiorlaan 42
1930 Zaventem
Belgium
Email: WVDV@starlab.net

Bill Vorn
5565 Clark
Montreal (Quebec)
Canada H2T 2V5
Email: d356644@er.uqam.ca

Eric Werner
Associate Professor of Computer Science
Computer Science
College of Arts and Sciences
Florida Gulf Coast University
19501 Treeline Avenue South
Fort Myers, Florida 33965
USA
Email: ewerner@fgcu.edu

Introduction

Kerstin Dautenhahn

This book discusses design issues of social agent technology with the perspective of human cognition. The book is written for an interdisciplinary audience, and each chapter is a self-contained contribution shedding light on a particular aspect of socially intelligent agents (SIA). The reader is not required to have a strong background in computer science, social science, or philosophy. This book is therefore distinguished from books in the field of computer science which are introducing technical issues of agent technology, and different from books in social psychology on the impact on and influence of human cognition on technology.

The idea to edit a book which combines aspects of human cognition with social agent technology arose from the AAAI Fall Symposium "Socially Intelligent Agents" (SIA) which was held in the Tang Center at MIT in Cambridge, Massachusetts, November 8-10, 1997. The Symposium brought together more than 50 researchers in the field which discussed different aspects of software and robotic agents. Since then, a number of workshops, symposia and publications developed these ideas further, and focussed on particular aspects of social agents (like trust, emotions, narrativity and many more). However, a reoccurring theme which crystallised from the discussions we had in Cambridge was that the notions of 'agent', 'social', and 'intelligence' strongly depend on the human-in-the-loop, namely the way humans perceive and interpret the world around them, including biological and artificial agents. Human cognition, and in particular social cognition, plays an important role for designers and users of agent technology, this was a central lesson. The message itself was not novel, and is familiar to us from discussions on second-order cybernetics and constructivism. However, to some of us it came as a surprise that an area which is said to be at the forefront of technological development discovers the important role of subjectivity and the

importance of the social and cultural background of designers and users of agent technology. Designing socially intelligent agents cannot be separated from the pursuit of fundamental questions about intelligence, agency, and how humans relate to technology.

Research on socially intelligent agents can have many faces:

1. Investigation of appropriate languages to model and implement agent systems.
2. Studies of agent architectures, their limitations and applicability.
3. Development of agent theories, which can drive experiments in order to build a science of agent research.

The goals of agent researchers can have many faces, too:

1. Agent research can aim at modelling biological agents, e.g. in social science research.
2. Agent research can aim at developing innovative technologies.

The concrete work of agent researchers can be very different, too, ranging from investigating formal agent models, studying computer simulations, building and experimenting with robots, designing and implementing user interfaces or virtual environments, studying and modelling animal behaviour, etc.

Research into socially intelligent agents is a strongly interdisciplinary enterprise and researchers in this field need to be on the one hand experts in their particular area but on the other hand show a strong interest to look beyond the boundaries of their field.

The chapters in this book present innovative and novel aspects to applications, methodology, theories, and/or implementations of agent systems. They have all in common two important ingredients:

1. They address the design, control, or development of artificial agents (software or hardware agents).
2. They relate this enterprise to human cognition, i.e. to the way humans perceive, act on, and interpret the world, in particular the social world.

The chapters in this book can be grouped differently, depending on different aspects which they address, and the interest and viewpoint of the reader. Putting the chapters in a sequential order was therefore a practical decision and not a natural one. In order to enable the reader to form a cognitive roadmap of the book the 16 chapters are in the following briefly described and embedded in the overall issue of human cognition and social agent technology. My personal view, opinions and thoughts on the issues which the

chapters address are put in italics in order to distinguish them from the description of the chapters themselves.

The chapters in this book were accepted after peer reviewing.

Chapters 1-3: Narrativity and Implementations of Story-Telling Systems

Chapter 1. Phoebe Sengers discusses insights from the field of narrative psychology with respect to designing agents from an Artificial Intelligence perspective. Humans seem to recognize and understand intentional behaviour in terms of narrative, which has important consequences for designing agents which should appear believable and intentional. In Senger's interpretation of the term 'narrative intelligence' agents can be designed which express intentionality as narratives, which allows in turn humans to perceive and explain the behaviour of these agents in the 'natural' way of narratives.

Thus, artificially intelligent agents which seem truely alive might need a 'natural' talent to tell stories.

The following two chapters describe two different design implementations (SAGE and ROSEBUD) which allow interactions between children and digital storytellers, mediated by interactive stuffed animals. This approach builds on the children's tendencies to share their personal stories with such soft interactive interfaces. Additionally, the child using these systems is at the same time audience and author, an approach different from the traditional passive role of users as consumers. The realization of these projects required to integrate different kinds of agents: a) system designers and their theoretical considerations which initialized the projects, b) synthetic agents, i.e. the design of software (digital storytellers) and digitally augmented physical (stuffed animals) agents, c) children as authors, interactors, and test persons to evaluate the systems.

In addition to an entertaining and educational aspects, interactive story-telling systems can potentially empower children to develop their self and inner world by authoring their own autobiography.

Chapter 2. Jennifer Williamson Glos discusses the cognitive, social, and psychological aspects of the ROSEBUD system. The implementation and the theoretical framework are described. 'Tangible media', like everyday 'keepsake objects' (e.g. stuffed animals), can merge the physical and the digital world and satisfy users' social and cognitive needs. Digital keepsakes also

have the potential to bridge the gender gap, according to Justine Cassell's notion of feminist software design in which authority and control is given to the users.

Chapter 3. Marina Umaschi Bers and Justine Cassell present the SAGE (Storytelling Agent Generation Environment) authoring environment for children to create their own wise storytellers which they can interact with. The chapter describes SAGE's design and implementation, and results from pilot studies with children. The chapter also discusses how the system assists children exploring story-telling and communication structures which helps them exploring their own identify, ways to present themselves to others, as well as taking the point of view of others.

Chapters 4 and 5: Socially Situated Software Agents

Chapter 4. Linda Cook, Tim Bickmore, Sara Bly, Elizabeth Churchill, Scott Prevost and Joseph W. Sullivan describe a system which allows users to create a personal representative which can, in a limited capacity, act on their behalf. Such a semi-autonomous agent needs to be socially competent, have a personality and a degree of similarity with the user. The history of the project, basic design considerations, as well as a prototype application, namely document avatars which express author's personal views about the content of a document, are presented. The project is discussed in the general context of personality and self-representatives in computer-mediated communication and work spaces.

Research into personal representatives bears the promise that we will actually be able to be at different places simultaneously, not only computationally (the original idea of agents which act on behalf of the user), but also visually and behaviourally. The task to integrate the parallel existences of multiple, similar but differently adapted personalities and appearances into a single autobiography and lifeline will remain the task of the biological human agent.

Chapter 5. Myles Bogner, Uma Ramamurthy and Stan Franklin describe a socially situated and 'conscious' software agent, Conscious Mattie. The authors base their notion of consciousness on the global workspace theory of consciousness proposed by Baars. The authors aim to build a software agent which can act as a seminar coordinator, communicating via email. Software architecture and functionality of CMattie are discussed. A particular focus is

given to the mechanisms which make the software agent 'conscious' and which allow conceptual learning. In addition to building an intelligent 'conscious' software agent, the authors implementation can provide testable hypotheses to neuroscientists and cognitive scientists.

Many theories of consciousness lack testable hypotheses, so that discussions often remain on the level of thought experiments with no or little relevance to technological developments. Instead, the design of socially situated agents can generate data which theories on human intelligence and cognition can be evaluated against. However, if this synthetic approach will create agents which are genuinely conscious, then might observing, dissecting and analysing them become as controversial as it is today for monkeys and humans? Consciousness in animals is often discussed with respect to whether animals can recognise themselves in a mirror, e.g. remove a spot placed during sleep on their forehead. How can an appropriate mirror-test for CMattie look like?

Chapters 6-10: Architectures for Socially Intelligent Agents

Chapter 6. Dolores Cañamero and Walter Van de Velde discuss how emotions ground perception and social interaction. A synthetic physiology is modelled which involves motivations, drives, and emotional states for an autonomous agent. The authors suggest to apply this model to a scenario on co-habited mixed reality which involves large heterogeneous societies decomposed in loosely coupled communities, and developing technologies that could enhance social processes within such societies. More specifically, the role of emotions is discussed with respect to the COMRIS (Co-Habited Mixed Reality Information Spaces) project which studies hybrid societies of humans and software agents.

Chapter 7. Aaron Sloman discusses architectural requirements for human-like agents and explains why only an information processing machine can love. Machines, made of software or hardware, might be able to show different kinds of love, similar to the same way humans do. In several stages different types of agent architectures and their information processing capacities are presented and discussed with respect to emotional processes. Architectural layers comprise mechanisms for reactive, deliberative, self-management, and non-semantic control. Research into layers of information processing can lead to different architectures which can be used to design artificial agents,

and possibly better understand biological agents.

The fact that computer programs can be good or even better chess players than humans might be frustrating (in particular to chess players) but it still gives us the chance to say that there is more to being human than problem-solving, namely the inner, phenomenological world, and the realm of emotions. Models of emotions can give rise to efficient adaptive control of artificial agents. Artificial agents as-we-know-them can simulate emotions, express emotions, appear emotional, and exploit the human tendency to anthropomorphize and attribute intentions and mental attributes. But what if artificial agents as-they-could-be might genuinely have emotions, if we have scientifically no argument left for defending our human (animal) emotions as superior to emotions of artificial agents?

Chapter 8. Ruth Aylett and David Barnes contrast two alternative control approaches in Artificial Intelligence: explicit reasoning between alternatives (the reflective approach) and situated dynamic agent-environment interactions (the behavioural approach). In order to combine the advantages of both approaches, a hybrid multi-agent architecture is studied in the project MACTA (Multiple Automata for Complex Task Achievement) and applied to cooperative behaviour of two mobile robots. General issues of intelligence and 'consciousness' of such a heterogeneous multi-agent model are discussed and compared with the results from the robot experiments.

Social insect societies show impressive examples of behavioural complexity which presumably do not require a 'reflective' component. Incrementally evolving robotic systems of increasing complexity, mimicking natural evolution, is currently challenged in the field evolutionary robotics. However, how to scale up from insect to mammal-like behaviour of robots still has to be shown. Then, hybrid approaches become promising control architectures for intelligent robots solving complex tasks where reasoning and planning processes are useful for the agents. In any case, a control architecture for a robot has to consider the particular physical nature, embodiment, of a robot and its environment.

There might be many different architectures for agents to show behaviour a bit like us (being emotional, capable of intelligent problem-solving), and some of these architectures can be suitable for computationally (software) and physically (hardware) embodied agents. The choice of architecture depends on many factors, e.g. whether evolutionary relevance of the architecture matters (is the architecture supposed to be a model of animal cognition).

Moreover, the choice depends on how it a) empowers the agent to act and interact autonomously in the world, and b) how it empowers the human (as a designer, experimenter, or user) to control and understand the agent's behaviour.

Chapter 9. Maria Miceli and Cristiano Castelfranchi address the cognitive and social role of evaluations in problem solving and decision making. Evaluations and values are presented as particular mental attitudes which are important for a goal-pursuing agent which is interacting with other agents. Evaluating the usefulness of an object with regard to a goal can be achieved by different kinds of evaluative processes, namely performance, analogy, classification, or by standard. Evaluations and emotional appraisal are not identical, but have a similar function, namely controlling the acquisition of adequate means for one's goals. In a social context, the ability to evaluate oneself, others, and the relationship of oneself to others plays a crucial role.

Human societies are complex structures where relationships and social interactions between individuals are neither fully engineered (by nature) nor entirely enslaved by opportunities provided by the environment. The study of cognitive processes which allow to model the dynamic Web of human social relationships and the way individual humans regulate and control their relationships to the social and non-social world is therefore an important step towards understanding human (social) cognition.

Chapter 10. Eric Werner discusses how an individual autonomous agent is integrated into a social system and becomes a social being. The focus is hereby on internal, psychological processes inside an agent. The development of a social self, which gives the agent a social competence, is discussed by linking formal concepts from multi-agent research with ideas as internalization and socialization which originate in psychoanalysis. This is a speculative first attempt towards a formal computational theory for the development of the social self. This line of research could be applied to understanding psychodynamic phenomena in both human and artificial agents.

Human agents are not born with fully developed social competence, so modelling the ontogeny of human social behaviour can help to have agents develop their own social competence which is adapted to the social and non-social environment where they are interacting in.

Chapters 11-16: Design Issues for Socially Situated Interactive Systems

Chapter 11. Christopher Landauer and Kirstie L. Bellman address the development of computationally embodied, complex and autonomous agents in social environments. In their view intelligent agency requires the creation of a) 'ecological niches' for computational agents, b) an artificial embodiment for these agents, c) agents with social behaviours, d) the capability for growth and adaptivity of behaviours. For the latter, reflective processes like prediction, planning and discovery are necessary. The Wrapping approach has been developed in order to construct heterogeneous software and hardware environments for creating computationally reflective systems. MUVE's (multi-user virtual environments) are discussed as testbeds of this approach.

There is more to embodiment than having a physical body. Being explicit about properties of embodiment and analysing what they can mean to computational or hardware agents is scientifically more constructive than arguing about the real or artificial nature of agents and environments.

Chapter 12. Leonard N. Foner introduces JULIA, a MUD (Multi-User Dungeon) software agent. JULIA is a social agent that gives navigational and other assistance in a MUD. In this particular domain it is able to interact with people as if it were another person. For many users JULIA passed the Turing-Test, namely it can (via written natural language dialogues) successfully pretend to be human! Foner addresses general issues of human cognition, e.g. expectations and the tendency to anthropomorphize agents, and how they might help and hinder designing social agents. Although the computational technology behind JULIA, e.g. her discourse model, is relatively simple, the design effectively exploits specific characteristics of the domain for which the agent is designed.

On the one hand, social agent technology addresses models, architectures and mechanisms which control an agent and its interaction with other agents. On the other hand, the example of JULIA clearly shows that the success of a social agent design need not be reflected in the complexity and degree of sophistication of the agent model. Rather, a careful domain analysis can specify the niche and the interaction space to which the agent design can be adapted. Such designs are not universal, but can produce successful and efficient designs.

Chapter 13. Stephen Grand discusses the design of software agents that people enjoy to interact with and treat as living, socially functioning beings.

The Norns and their relatives, inhabitants of the computer game Creatures, have populated millions of computers worldwide, and have transformed the general view on the nature of 'pets'. The design follows a deeply biological approach and behaviours are generated through a structural, rather than a procedural paradigm. Biological building blocks which the agent design is based on comprise an artificial neuronal brain, an artificial chemistry, and an artificial genetics. The building blocks allow to engineer emergent life-like phenomena, although the phenomena themselves are not explicitly represented.

People relate to computer programs and agents in a similar way as they relate to humans. The personality and presence of software pets can become as real as the personality of biological pets. The issue when increasingly complex artificial beings cross the threshold from models or simulations of life to truely living and conscious beings remains to be shown.

Chapter 14. Bill Vorn argues that real or virtual agents are not only an industrial, scientific or artistic medium, but also a medium of communication which can carry informational content (messages, ideas, cultural values). He addresses this issue from two different angles. Firstly, he introduces his work as a designer of artwork, namely robotic interactive installations. Secondly, he presents a survey study on Artificial Life as a communication phenomenon which indicates that people attribute meaning and signification to Artificial Life agents, but do not necessarily consider Artificial Life as a media.

The artifacts we produce might be able to speak, but they can also speak for themselves, representing the social and cultural context of their creation. The synthesis of an artifact is a creative act and as such transports meaning and content. Designs of social agents transmit a notion of sociality. Imagine that beauty and dignity are the messages of biological designs, what messages do we want our artifacts to convey?

Chapter 15. Simon Penny discusses agent building from the perspective of the visual arts. He argues that in the special context of agent design aspects of the artistic methodology can complement conventional scientific methodologies. These contributions from artistic practise are more specifically: the embodied experience of the user, the cultural specificity of any representational act, and the critical importance of the interface. Meaning is not engineered, it is established in the cultural environment of the interaction between user and agent. Penny's arguments are supported by descriptions of his own work, namely an autonomous interactive robot, an interactive sound installa-

feelings, and perspectives on the world.

How can we build creatures that are not just smart but visibly alive in this way? In this paper I will try to provide some answers by turning the question inside out. In order to build creatures that seem intentionally alive, we can try to understand how *human beings* identify and interpret intentional behavior. If we understand what properties enable people to understand behavior as intentional, we may be able to generate a similar style of behavior in artificial agents. Such agents, being engineered to be intentionally understandable, are more likely to really seem alive than an agent optimized for formal intelligence or for use as a goal-oriented tool.[1]

Narrative psychology suggests that people understand the behavior of living agents by structuring visible activity into narrative (Bruner 1990). That is, people understand and interpret intentional behavior by organizing it into a kind of story. If this is the case, then our agents may appear more intentional if we build them so that their behavior provides the cues to be understandable as narrative. Here, I will describe the fundamental principles of narrative in order to explain (1) how current agent construction techniques actually undermine the appearance of intentionality and (2) how narrative principles can be applied in agent design to support the user in understanding the agent as an intentional being.

2. Principles of Narrative Psychology (or How We (Sometimes) Make Sense of Creatures)

Artificial Intelligence attempts to generate intentional creatures by setting up a correspondence between biological, living beings and automatic processes of the kind that can run on computers. That is, AI agents should ideally be understandable both as well-specified physical objects and as sentient creatures. But it turns out that there is a deep tension between these two views on agents. This is because human understanding of the behavior of humans and other conscious beings differs in important ways from the way we understand the behavior of such physical objects as toasters. Identifying the distinction between these two styles of comprehension is essential for discovering how to build creatures that are understandable not just as helpful tools but as living beings.

The way people understand meaningful human activity is the subject of

narrative psychology, an area of study developed by Jerome Bruner (1986, 1990). Narrative psychology shows that, whereas people tend to understand inanimate objects in terms of cause-effect rules and by using logical reasoning, intentional behavior is made comprehensible, not by figuring out its physical laws, but by structuring it into narrative or 'stories.' This structure is not simply observed in the person's activity; we generate it through a sophisticated process of interpretation. This interpretation involves such aspects as finding relations between what the person does from moment to moment, speculating about how the person thinks and feels about his or her activity, and understanding how the person's behavior relates to his or her physical, social, and behavioral context.

Even non-experts can effortlessly create sophisticated interpretations of minimal behavioral and verbal cues. In fact, such interpretation is so natural to us that when the cues to create narrative are missing, people spend substantial time, effort, and creativity trying to come up with possible explanations. This process can be seen in action when users try to understand our currently relatively incomprehensible agents!

This sometimes breathtaking ability — and compulsion — of the user to understand behavior by constructing narrative may provide the key to building agents that truly appear alive. If humans understand intentional behavior by organizing it into narrative, then our agents will be more *intentionally comprehensible* if they provide narrative cues. That is, rather than simply presenting intelligent actions, agents should give visible cues that support users in their ongoing mission to generate narrative explanation of an agent's activity. We can do this by organizing our agents so that their behavior provides the visible markers of narrative. The remainder of this paper presents the properties of narrative and explains how they apply to agent construction.

3. Prolegomena to a Future Narrative Intelligence

There has recently been a groundswell of interest in narrative in AI and human-computer interaction (HCI). Narrative techniques have been used for applications from automatic camera control for interactive fiction (Galyean 1995) to story generation (Elliott et al. 1998). Abbe Don (1990) and Brenda Laurel (1986, 1991) argue that, since humans understand their experiences in terms of narrative, computer interfaces will be more understandable if they are

organized as narrative. Similarly, Kerstin Dautenhahn and Chrystopher Nehaniv (1998) argue that robots may be able to use narrative in the form of autobiography to understand both themselves and each other.

Michael Travers and Marc Davis developed the term Narrative Intelligence in the context of an informal working group at the MIT Media Lab to describe this conjunction of narrative and Artificial Intelligence. David Blair and Tom Meyer (1997) use the same term to refer to the *human* ability to organize information into narrative. Here, I want to suggest that Narrative Intelligence can be understood as the confluence of these two uses: *that artificial agents can be designed to produce narratively comprehensible behavior by structuring their visible activity in ways that make it easy for humans to create narrative explanations of them.*

In order to do this, we need to have a clear understanding of how narrative works. Fortunately, the properties of narrative have been extensively studied by humanists. Bruner (1991) nonexhaustively lists the following properties:

- *Narrative Diachronicity*: Narratives do not focus on events on a moment-by-moment basis, but on how they relate over time.
- *Particularity:* Narratives are about particular individuals and particular events.
- *Intentional State Entailment*: When people are acting in a narrative, the important part is not what the people do, but how they think and feel about what they do.
- *Hermeneutic Composability*: Just as a narrative comes to life from the actions of which it is composed, those actions are understood with respect to how they fit into the narrative as a whole. Neither can be understood completely without the other. Hence, understanding narrative requires interpretation in a gradual and dialectical process of understanding.
- *Canonicity and Breach*: Narrative makes its point when expectations are breached. There is a tension in narrative between what we expect to happen and what actually happens.
- *Genericness*: Narratives are understood with respect to genre expectations, which we pick up from our culture.
- *Referentiality*: Narratives are not about finding the absolute truth of a situation; they are about putting events into an order that feels right.
- *Normativeness*: Narratives depend strongly on the audience's conventional expectations about plot and behavior.
- *Context Sensitivity and Negotiability*: Narrative is not 'in' the thing being

understood; it is generated through a complex negotiation between reader and text.

– *Narrative Accrual*: Multiple narratives combine to form, not one coherent story, but a tradition or culture.

While these properties are not meant to be the final story on narrative, they stake out the narrative landscape. Taking narrative agents seriously means understanding how these properties can influence agent design. It will turn out that current AI techniques, which largely inherit their methodology from the sciences and engineering, often undermine or contradict the more humanist properties of narrative. Here, I will explain problems with current agent-building techniques, techniques already in use that are more amenable to narrative, and potential practices that could be more friendly to the goal of meaningful Narrative Intelligence.

One note of caution: the goal here is to interpret the properties of narrative with respect to agent-building. This interpretation is itself narrative. Since, as we will see below, the nature of narrative truth is different from that of scientific factuality, this essay should not be read in the typically scientific sense of stating the absolute truth about how narrative informs AI. Rather, I will look at the properties of narrative in the context of current AI research, looking for insights that might help us to understand what we are doing better and suggest (rather than insist on) new directions. My conclusions are based on my particular human perspective, as a builder of believable agents in Joseph Bates' Oz Project with a strong interest and training in the cultural aspects of Artificial Intelligence.

3.1 *Narrative diachronicity*

The most basic property of narrative is its diachronicity: a narrative relates events *over time*. Events are not understood in terms of their moment-by-moment significance, but in terms of how they relate to one another as events unfold. For example, if Fred has an argument and then kicks the cat, we tend to infer that the cat-kicking is not a random event, but a result of his frustration at the argument. When people observe agents, they do not just care about what the agent is doing; they want to understand the relations between the agent's actions at various points in time. These perceived relations play an important role in how an agent's subsequent actions are understood. This means that, to be properly understood, it is important for agents to express their actions so

that their intended relationships are clear.

However, it is currently fashionable to design behavior-based autonomous agents using action-selection, an agent-building technique that ignores the diachronic structure of behavior. Action-selection algorithms work by continuously redeciding the best action the agent can take in order to fulfill its goals (Maes 1989a). Because action-selection involves constantly redeciding the agent's actions based on what is currently optimal, behavior-based agents often display a kind of "schizophrenia" (Sengers 1996b). By schizophrenia I mean that they jump from behavior to behavior, without any kind of common thread that structures these behaviors into understandable sequences. Schizophrenic agents undermine the appearance of intentionality because agent action seems to be organized arbitrarily over time, or, at maximum, in terms of automatic stimulus-response.[2]

More generally, expressing the relationships between behaviors is not well supported in most behavior-based systems (a complaint also raised in (Neal Reilly 1996)). While these architectures do provide support for clear, expressive *individual* behaviors, they have problems when it comes to expressing relations *between* behaviors. This is because a typical behavior-based system (e.g. Blumberg 1994; Brooks 1986a; Maes 1991) treats each behavior separately; behaviors should refer as little as possible to other behaviors. Because of this design choice, a behavior, when turned on, does not know why it is turned on, who was turned on before it, or even who else is on at the same time. It knows only that its preconditions must have been met, but it does not know what other behaviors are possible and why it was chosen instead of them. In most behavior-based architectures, behaviors simply do not know enough about other behaviors to be able to express their interrelationships to the user.

In this light, classical AI would seem to have an advantage over alternative AI, since it is explicitly interested in generating structured behavior through such mechanisms as scripts and hierarchical plans. However, classical AI runs into similar trouble with *its* modular boundaries, which occur not between behaviors but between the agent's functionalities. For example, the agent may say words it cannot understand, or clearly perceive things that then have no influence on what the agent decides to do.

Fundamentally, agent-building techniques from Marvin Minsky's Society of Mind (1988) to standard behavior-based agent-building (Maes 1991) to the decomposition of classical agents into, for example, a planner, a natural

language system, and perception (Vere and Bickmore 1990) are all based on divide-and-conquer approaches to agenthood. Being good computer scientists, one of the goals of AI researchers is to come up with modular solutions that are easy to engineer. While some amount of atomization is necessary to build an engineered system, narrative intentionality is undermined when the parts of the agent are designed so separately that they are visibly disjoint in the behavior of the agent. Schizophrenia is an example of this problem, since when behaviors are designed separately the agent's overall activity reduces to a seemingly pointless jumping around between behaviors. Bryan Loyall similarly points out that visible module boundaries destroy the appearance of aliveness in believable agents (Loyall 1997).

The end result is that the seductive goal of the plug-n-play agent — built from the simple composition of arbitrary parts — may be deeply incompatible with intentionality. Architectures like that of Steels (1994) which design behaviors in a deeply intertwined way, make the agent design process more difficult, but may have a better shot at generating the complexity and non-modularity of organic behavior. Less drastic solutions may involve the use of transition sequences to relate and smooth over the breaks between separately designed behaviors (Stone 1996) (Sengers 1998b). I use this strategy elsewhere as a cornerstone for the Expressivator, an architecture for Narrative Intelligence (1998a).

3.2 *Particularity*

Narratives are not simply abstract events; they are always particular. "Boy-meets-girl, boy-loses-girl" is not a narrative; it is the structure for a narrative, which must always involve a particular boy, a particular girl, a particular way of meeting, a particular way of losing. These details bring the story to life. However, details do not by themselves make a narrative either; the abstract structure into which the details can be ordered brings meaning to the details themselves. A narrative must be understood in terms of tension between the particular details and the abstract categories they refer to; without either of these, it is meaningless.

This same tension between the abstract and the particular can be found in agent architectures. Agent designers tend to think about what the agent is doing in terms of abstract categories: the agent is eating, hunting, sleeping, etc. However, users who are interacting with the agent do not see the abstract

categories; they only see the physical movements in which the agent engages. The challenge for the designer is to make the agent so that the user can (1) recognize the particular details of the agent's actions and (2) generalize to the abstract categories of behavior, goal, or emotion that motivated those details. Only with a full understanding at both the particular and the abstract levels will the user be likely to see the creature as the living being the designer is trying to create.

But AI researchers are hampered in this full elucidation of the dialectical relationship between the particular and the abstract by the valorization of the abstract in computer science. In AI we tend to think of the agent's behaviors or plans as what the agent is 'really' doing, with the particular details of movement being a pesky detail to be worked out later. In fact, most designers of agents do not concern themselves with the actual working out of the details of movement or action at all. Instead, they stop at the abstract level of behavior selection, reducing the full complexity of physical behavior to an enumeration of behavior names. Maes (1989b), for example, uses abstract atomic actions such as "pick-up-sander."

Similarly, the Oz Project's first major virtual creature, Lyotard, was a text-based virtual cat (Bates et al. 1992). Because Lyotard lived in a text environment, his behaviors were also text and therefore high level: "Lyotard jumps in your lap," "Lyotard eats a sardine," "Lyotard bites you." Because we were using text, we did not need to specify action at a more detailed level. We did not have to specify, for example, how Lyotard moved his legs in order to jump in your lap.

Lyotard's successors, the Woggles (Loyall and Bates 1993), on the other hand, were graphically represented. As a consequence, we were forced to specifically define every low-level action an agent took as part of a behavior. The effort that specification took meant that we spent less time on the Woggles' brains, and as a consequence the Woggles are not as smart as Lyotard. But — surprisingly to us — the Woggles also have much greater affective power than Lyotard. People find the Woggles simply more convincingly alive than the text cat, despite the fact that Lyotard is superior from an AI point of view. This is probably in part because we were forced to define a particular body, particular movements, and all those pesky particularities we AI researchers would rather avoid.[3]

If we look at animation (e.g. Thomas and Johnston 1981), the valorization tends to run to the other extreme: the particular is seen as the most

essential. Animators tend to think mostly at the level of surface movement; this movement may be interpretable as a behavior, as evidence of the character's emotions, as revealing the character's motivations, or as any of a host of things or nothing at all. Animators make the point that any character is of necessity deeply particular, including all the details of movement, the structure of the body, and quirks of behavior. The abstract comes as an afterthought. Certainly, animators make use of a background idea of plot, emotion, and abstract ideas of what the character is doing, but this is not the level at which most of animators' thinking takes place.

Loyall (1997) points out that this focus on the particular is also essential to the creation of effective believable agents. A focus on particularity by itself, though, is not adequate for creating artificial agents. Agents are expected to interact autonomously with the user over time. In order to build such autonomous systems, we need to have some idea of how to structure the agent so that it can recognize situations and react appropriately. Because we do not know every detail of what will happen to the agent, this structure necessarily involves abstract concepts in such aspects as the modules of the agent, the classification of situations according to appropriate responses, abstract behaviors, emotions, goals, and so on.[4] We must design agents, at least partially, at an abstract level.

In order to build agents that effectively communicate through narrative, AI researchers will need to balance their ability to think at the abstract level with a new-found interest in the particular details their system produces, an approach that seems to be gaining in popularity (Frank et al. 1997). Narrative Intelligence is only possible with a deep-felt respect for the complex relationship between the abstract categories that structure an agent and the physical details that allow those categories to be embodied, to be read, and to become meaningful to the user.

3.3 *Intentional state entailment*

Suppose you hear the following:

> A man sees the light is out. He kills himself.

Is this a story? Not yet. You don't understand it.

After endless questions, you find out that the man was responsible for a light house. During the night, a ship ran aground off shore. When the man sees

that the light house light is out, he realizes that he is responsible for the shipwreck. Feeling horribly guilty, he sees no choice but to kill himself. Now that we know what the man was thinking, we have a story.

In a narrative, what actually happens matters less than what the actors feel or think about what has happened. Fundamentally, people want to know not just what happened but *why* it happened. This does not mean the causes of an event in terms of physical laws or stimulus-response reactions, but the reasons an actor freely chose to do what s/he did. The narrative is made sense of with respect to the thoughts and feelings of the people involved in its events.

This means that when people watch autonomous agents, they are not just interested in what the agent does. They want to know how the agent thinks and feels about the world around it. Instead of knowing only what the agent has chosen to do, they want to know *why* the agent has chosen to do it.

But in many autonomous agent architectures, the reasons for the decisions the agent makes are part of the implicit architecture of the agent and therefore not directly expressible to the user. Bruce Blumberg's Hamsterdam architecture, for example, represents the appropriateness of each currently possible behavior as a number; at every time step the behavior with the highest number is chosen (Blumberg 1996). With this system, the reasons for behavioral choice are reduced to selecting the highest number; the actual reason that behavior is the best is implicit in the set of equations used to calculate the number. The agent simply does not have access to the information necessary to express why it is doing what it does.

Instead of this emphasis on *selecting* the right action, Tom Porter (1997) suggests the strategy of *expressing* the reasons an agent does an action and the emotions and thoughts that underly its activity. This means organizing the agent architecture so that reasons for behavioral change are explicit and continuously expressed. By showing not only what the agent does, but why the agent does it, people may have an easier time understanding what the agent is thinking and doing in general.

A deeper problem with current architectures is that ethologically-based models such as (Blumberg 1996) presuppose that most of what an agent does is basically stimulus-response. But when we build agents that embody these theories, they often work through stimulus-response or straightforward cause-effect. This automaticity then carries forward into the quality of our agent's behavior — the agent seems nothing more than it is, an unthinking automaton.

More generally, as scientists, we are not interested in the vagaries of free

will; we want to develop clearly-specified rules to explain why animals do what they do when they do it. In particular, in order to embody our ideas of agenthood in an automatically-running computational architecture, we must intentionally adopt what Daniel Dennett (1987) might call a 'non-intentional stance' — there is no algorithm for "and then the agent should do whatever it feels like doing." We therefore tend to develop theories of behavior that are fundamentally mechanistic.

But these mechanistic theories of agenthood often lead to mechanistic qualities of behavior in our generated agents. As a consequence, agents are not only non-intentional for us; they are also often reduced to physical objects in the eyes of the user. Narrative Intelligence requires agents that at least appear to be thinking about what they are doing and then making deliberate decisions according to their own feelings and thoughts, rather than simply reacting mindlessly to what goes on around them. We may be automatic; but we should not appear so.

3.4 *Hermeneutic composability*

Narrative is understood as a type of communication between an author and an audience. In order to understand this communication, the audience needs to go through a process of interpretation. At the most basic level, the audience needs to be able to identify the atomic components or events of the narrative. But this is just the beginning; the audience then interprets the events not in and of themselves but with respect to their overall context in the story. Once the story is understood, the events are re-identified and re-understood in terms of how they make sense in the story as a whole. In essence, this is a complex and circular process: the story only comes into being because of the events that happen, but the events are always related back to the story as a whole.

This property of narrative is another nail in the coffin of the dream of plug-n-play agents. If users continuously re-interpret the actions of the agent according to their understanding of everything the agent has done so far, then agent-builders who design the parts of their agents completely separately are going to end up misleading the user, who is trying to understand them dialectically.

More fundamentally, the deep and complex interrelationships between the things creatures do over time is part of what makes them come alive, so much so that when there are deep splits between the parts of a person — for

example, they act very happy when they talk about very sad things — we consider them mentally ill. This kind of deep consistency across parts is very difficult to engineer in artificial systems, since we do not have methodologies for engineering wholistically. In alternative AI, it is currently fashionable to believe that these deep interrelationships may come about *emergently* from separately designed pieces; whether this is wishful thinking or the foundation for a novel form of wholistic design is not yet clear. It may be that the best we can do is the surface impression of wholism; whether that will be enough remains to be seen.

3.5 *Canonicity and breach*

A story only has a point when things do not go the way they should. "I went to the grocery store today" is not a story; but it is the beginning of a story when I go on to say "and you'll never believe who I ran into there." There is no point to telling a story where everything goes as expected; there should be some problem to be resolved, some unusual situation, some difficulty, someone behaving unexpectedly.... Of course, these deviations from the norm may themselves be highly scripted ("boy-meets-girl, boy-loses-girl, boy-wins-girl-back" being a canonical example).

It may be, then, that the impression of intentionality can be enhanced by making the agent do something unexpected. Terrel Miedaner's short story "The Soul of the Mark III Beast" (1981) revolves around just such an incident. In this story, a researcher has built an artificially intelligent robot, but one of his friends refuses to believe that a robot could be sentient. This continues until he hands her a hammer and tells her to destroy the robot. Instead of simply breaking down — the friend's canonical expectation — the robot makes sounds and movements that appear to show pain and fear of death. This shakes the friend so much that she starts to wonder if the robot is alive, after all. Watching the robot visibly grapple with its end, the friend is led to sympathy, which in turn leads her to see the robot as sentient.

More generally, people come to agents with certain expectations, expectations that are again modified by what they see the agent do. The appearance of intentionality is greatly enhanced when those expectations are not enough to explain what the agent is doing. That is, the agent should not be entirely predictable, either at the level of its physical actions or at the level of its overall behavioral decisions. Characters in a Harlequin romance — who

inevitably fall in love with the man they hate the most (James 1998) — have nowhere near the level of 3-dimensionality of the complex and quirky characters of a Solzhenitsyn novel. Similarly, agents who always do the same thing in the same situation, whose actions and responses can be clearly mapped out ahead of time, will seem like the automatons they are, not like fascinating living creatures.

Making the creature do unexpected things may seem like a contradiction to one of the basic goals of Narrative Intelligence: making agent behavior more understandable. Stereotypicity may seem like a helpful step towards making agent behavior comprehensible. After all, if the agent always does the same thing for the same reasons in the same ways, the user will always know exactly what the agent is doing. But since users are very good at creating narrative, stereotyped actions bore the audience. In order to create compelling narrative, there needs to be some work for the reader to do as well. The agent designer needs to walk the line between providing enough cues to users that they can create a narrative, and making the narrative so easy to create that users are not even interested.

3.6 *Referentiality*

The 'truth' in stories bears little resemblance to scientific truth. The point of stories is not whether or not their facts correspond to reality, but whether or not the implicit reasoning and emotions of the characters feels right. A plausible narrative does not essentially refer to actual facts in the real world, but creates its own kind of narrative world that must stand up to its own, subjective tests of realism.

Similarly, extensive critiques have been made in AI about the problem of trying to create and maintain an objective world model (Agre 1997). Having the agent keep track of the absolute identity and state of objects in the external world is not only difficult, it is actually unhelpful. This is because in many situations the absolute identity of an object does not matter; all that matters is how the agent wants to or could use the object. As a substitute, Philip Agre has introduced the notion of deictic representation, where agents keep track of what is going on, not in any kind of absolute sense, but purely with respect to the agent's current viewpoint and goals (Agre 1988).

While understanding the power of subjectivity for agents, AI in general has been more reluctant to do away with the goal of objectivity for agent

researchers. AI generally sees itself for better or for worse as a science, and therefore valorizes reproducibility, testability, and objective measures of success. For many, intelligence is or should be a natural phenomenon, independent of the observer, and reproducible in an objective manner. Intelligence is not about appearance, but about what the agent 'actually' does. This reveals itself in the oft-repeated insistence that agents should not just appear but *be* 'really' alive or 'really' intelligent — anything else is considered illusionary, nothing more than a bag of tricks.

This 'real' essence of the agent is usually identified with its internal code — which is also, conveniently enough, the AI researcher's view of the agent. As a consequence, the impression the agent makes on the user is often considered less real, and by extension, less important. This identification of the internal code of the agent as what the agent really *is* — with the impression on the user a pale reflection of this actual essence — has an unexpected consequence: it means that the subjective interpretation of the audience is devalued and ignored. The result is agents that are unengaging, incoherent, or simply incomprehensible.

This does not mean the AI community is idiotic. Most AI researchers simply have a scientific background, which means they do not have training in subjective research. But the accent on AI as a science, with the goals and standards of the natural sciences, may lose for us some of what makes narrative powerful. I do not believe that 'life' in the sense of intentionality will be something that can be rigorously, empirically tested in any but the most superficial sense. Rather, generating creatures that are truly alive will probably mean tapping into the arts, humanities, and theology, which have spent centuries understanding what it means to be alive in a meaningful way. While intelligent tools may be built in a rigorous manner, insisting on this rigor when building our 'robot friends' may be shooting ourselves in the foot.

3.7 *Genericness*

Culturally supplied genres provide the context within which audiences can interpret stories. Knowing that a story is intended to be a romance, a mystery, or a thriller gives the reader a set of expectations that strongly constrain the way in which the story will be understood. These genre expectations apply just as well to our interpretations of everyday experience. The Gulf War, for example, can be understood as a heroic and largely victimless crusade to

restore Kuwait to its rightful government or as a pointless and bloody war undertaken to support American financial interests, depending on the typical genre leanings of one's political philosophy.[5]

These genres within which we make sense of the world around us are something we largely inherit from the culture or society we inhabit. This means at its most basic that different kinds of agent behavior make sense in different cultures. For example, I once saw a Fujitsu demo of 'mushroom people' who would, among other things, dance in time to the user's baton. In this demo, the user went on swinging the baton for hours, making the mushroom people angrier and angrier. Finally, it was the middle of the night, and the mushroom people were exhausted, obviously livid — and still dancing. I thought this behavior was completely implausible. "Why on earth are they still dancing? They should just leave!" I was told, "But in Japan, that would be rude!" My American behavioral genre expectations told me that this behavior was unnatural and wrong — but in Japan the same behavior is correct.

Since cultural expectations form the background within which agent behavior is understood, the design of intentionally comprehensible agents needs to take these cultural expectations into account. In contrast, the current practice of building agents tends not to consider the specific context in which the agent will be used. Patricia O'Neill Brown (1997) points out that this is likely to lead to agents that are misleading or even useless. This means an understanding of the sociocultural environment in which an agent will be inserted is one important part of the agent design process. In fact, O'Neill Brown goes one step further: not only does cultural baggage affect the way agents *should* be designed, it *already* affects the way agents are designed. That is, the way designers think of agents has a strong influence on the way we build them to start out with.

This tension can even be seen within American culture (Sengers 1994). In particular, the American tradition of AI has included two competing visions of what it means to be an agent. Classical AI on the one hand tends to favor representational, deliberative, rational, cognitive agents. In contrast, alternative AI tends to argue for nonrepresentational, reactive, situated, and embodied agents.

From within AI, these two conceptions of agents can seem to stem purely from technical imperatives. With a broader perspective, they can be traced back to the culture in which AI is situated, which has a number of different traditions of conceptualizing what it means to be human. In the West, human

beings have traditionally been thought of through what cultural theorists call the Enlightenment model of consciousness: the mind is separated from the body, it is or should be fundamentally rational, and cognition divorced from emotion is the important part of experience. This form of agency is in many ways fundamentally equivalent to the notion of agent proposed by classical AI. At the same time, in the last 30 or 40 years this view of humanity has been challenged by the 'schizophrenic' model of consciousness (see e.g. Massumi 1992).[6] This model considers people to be immersed in and to some extent defined by their situation, the mind and the body to be inescapably interlinked, and the experience of being a person to consist of a number of conflicting drives that work with and against each other to generate behavior. Alternative AI is clearly inspired by this notion of being human.

The conclusion from this and similar analyses of the metaphors behind AI technology (e.g. (Wise 1998)) is that *AI research itself is based on ideas of agenthood we knowingly or unknowingly import from our culture.* Given that this is the case, our best bet for harnessing the power of culture so it works for AI instead of against it is what Agre calls a critical technical practice: the development of a level of self-reflective understanding by AI researchers of the relationship between the research they do and culture and society as a whole (Agre 1997).

3.8 *Normativeness*

Previously, we saw that a story only has a point when things do not go as expected, and that agents should similarly be designed so that their actions are not completely predictable. But there is a flip side to this insight: since the point of a story is based on a breach of conventional expectations, narratives are strongly based on the conventions that the audience brings to the story. That is, while breaking conventions, they still depend on those same conventions to be understood and valued by the audience.

Intentional agents, then, cannot be *entirely* unpredictable. They play on a tension between what we expect and what we do not. There needs to be enough familiar structure to the agent that we see it as someone like us; it is only against this background of fulfilled expectations that breached expectation comes to make sense.

3.9 *Context sensitivity and negotiability*

Rather than being presented to the reader as a *fait accompli*, narrative is constructed in a complex interchange between the reader and the text. Narrative is assimilated by the reader based on that person's experiences, cultural background, genre expectations, assumptions about the author's intentions, and so on. The same events may be interpreted quite differently by different people, or by the same person in different situations.

In building narrative agents, on the other hand, the most straightforward strategy is context-free: (1) choose the default narrative you want to get across; (2) do your best to make sure the audience has understood exactly what you wanted to say. The flaw in this strategy is that narrative is not one size fits all. It is not simply presented and then absorbed; rather, it is constructed by the user. In assimilating narrative, users relate the narrative to their own lived experience, organizing and understanding it with respect to things that have happened to them, their generic and conventional expectations, and their patterns of being. Narrative is the interface between communication and life; through narrative a story becomes a part of someone's existence.

This means the 'preformed narrative' that comes in a box regardless of the audience's interests or wishes is throwing away one of the greatest strengths of narrative: the ability to make a set of facts or events come to life in a meaningful way *for the user* — in a way that may be totally different from what someone else would see. Rather than providing narrative in a prepackaged way, it may be more advantageous to provide the *cues* for narrative, the building blocks out of which each user can build his or her unique understanding.

And if narrative is not the same for everyone, then narrative agents should not be, either. If narrative is fundamentally user-dependent, then inducing narrative effectively means having some ideas about the expected audience's store of experience and typical ways of understanding. Just as the author of a novel may have a typical reader in mind, the designer of an agent needs to remember and write for the people who will use that agent, relating the agent's projected experiences to the lived experience of the desired audience.

And just as the author of a novel does not expect every possible reader to understand its point, the author of an agent does not necessarily need to be disappointed if only some people understand what the agent is about. The statistical testing of an agent's adequacy over user population may miss the point as much as using bestseller lists to determine the quality of novels. It

may be that making the point well with a few users is better, from the point of view of the designer, than making the point adequately with many users.

3.10 *Narrative accrual*

Generally speaking, narratives do not exist as point events. Rather, a set of narratives are linked over time, forming a culture or tradition. Legal cases accumulate, becoming the precedents that underly future rulings. Stories we tell about ourselves are linked together in a more-or-less coherent autobiography.

The mechanism by which narratives accrue is different from that of scientific fact. We do not find principles to derive the stories, or search for empirical facts in the stories to accept or reject according to a larger paradigm. Stories that contradict one another can coexist. The Bible, for example, first cheerfully recounts that, on the 7th day, God made man and woman at the same time; a little later, God makes man out of mud, and only makes woman after man is lonely (Various 1985). Similarly, we do not necessarily have a problem reconciling two stories, in one of which Fred is mean, and in the other he is nice. The process of reconciliation, by which narratives are joined to create something of larger meaning, is complex and subtle.

The ways in which stories are combined — forming, if not a larger story, at least a joint tradition — are not currently well understood. Once we have a better understanding of how this works, we could use these mechanisms in order to modulate the effects of our narrative agents as they move from episode to episode with the user. As Dautenhahn (1997) has suggested, agents are understood by constructing 'biographies' over the course of prolonged interaction. By investigating the mechanisms whereby the user constructs these biographies from the mini-narratives of each encounter, we stand a better chance of building our agent so that over time it makes an appropriate impression on the user.

4. Narrative Intelligence as a Research Tradition

Now that we have learned about the properties of narrative and what they might have to do with agents, we can return with a broader perspective to the questions raised in the introduction: why, as Rodney Brooks wonders, our agents simply do not seem alive in the way that natural creatures do. Having

looked at the nature of narrative in detail, I believe that at least part of the problem in AI research is its inheritance from scientific research traditions, through which — mostly unconsciously — the life in our creatures can be drained away. In this final section, I will first explain why an overreliance on science can undermine narrative, and then look at the implications of this problem for Narrative Intelligence as a field.

4.1 *Narrative and atomization*

In general, narrative involves understanding the wholistic relationships between things: the relationship between the different events in the story, the relationship between the events and how the actors feel about the events, the relationship between what the author tries to tell and the way in which the audience constructs what it hears, the relationship between the audience member and his or her cultural background, and so on. With layer upon layer of interdependency, this narrative view of the world can become extremely complex.

In contrast, the scientific worldview tends to value simplicity through black-boxing, a process I have elsewhere (1998a) termed *atomization*. Atomization is the process of splitting something that is continuous and not strictly definable into reasonably well defined, somewhat independent parts. In classical AI, for example, we atomize agents by dividing them into relatively independent functionalities, such as a planner, an executor, and a perception system. In alternative AI, we atomize agents by dividing them into relatively independent behaviors.

We do this for a good reason: atomization is a way of getting a handle on a complex phenomenon, a way of taking something incoherent, undefined, and messy and getting some kind of fix on it. It is only through atomization that we can understand something clearly enough to be able to engineer a working system of any complexity. In this sense, atomization is essential to AI.

But atomization as used in science is not a transparent methodology. In many ways, its properties are the exact opposite of those of narrative. This can be seen more concretely by inverting each of the properties of narrative:

1. *Structure over time*: Narrative structure is diachronic; it is about how events relate to one another. Atomistic structure is statistical. Patterns of events over time are simply correlated with one another.

2. *Essence*: Narrative is interested in particular events; it matters which person a story is about. Atomization is interested in finding salient properties so that events can be generalized as parts of a system; individual water molecules, for example, are not differentiated. Narrative sees events as essentially particular; atomization, as essentially abstract, with specific features seen as noise.
3. *Components*: Narrative is interested in events mainly in terms of how the actors involve understand and interpret them. Scientific atomization is interested in the facts that can be established independently of any one person's experience.
4. *Combination*: Narrative is wholistic; the act of bringing its components together changes the components themselves. In atomization, the combination of events is seen as the sum of the parts.
5. *Relation to expectation*: Narrative must contain elements that are unexpected; things cannot go as planned. In contrast, the goal of scientific atomization is to be able to predict and control with reasonable certainty the outcome of events.
6. *Referentiality*: Narrative is fundamentally subjective; it is about how different people come to interpret it in different situations. Scientific atomization is meant to be objective. Its laws hold in every situation, independent of context and interpretation.
7. *Dependence on culture*: Similarly, while narrative is largely dependent on culturally bound norms and expectations, scientific atomization is thought of as culturally universal, true for everyone.
8. *Audience judgment*: The audience must use its judgment for narrative to be realized; but audience judgment is considered to be unimportant for determining the truth of scientific atoms.
9. *Application*: The way in which narrative is used depends on context; atomic facts are meant to be absolute.
10. *Accrual*: Narratives combine to form a not necessarily coherent tradition. Atomic facts are combined by comparing them and finding a logical structure that subsumes them. Facts that are inconsistent are thrown away.

These aspects are summarized in Table 1.

Table 1. Relations between scientific (atomistic) and humanist (narrative) worldviews

Aspect	*Scientific Worldview*	*Humanist Worldview*
structure over time	statistical	diachronic
essence	abstract	particular
components	factual	experiential
combination	additive	wholistic
relation to expectation	predictable	creative
referentiality	objective	subjective
dependence on culture	culturally universal	culturally variable
audience judgment	unimportant	essential
application	absolute	context-sensitive
accrual	logical coherence	tradition

Clearly, these statements are too absolute. Not all scientific work is, for example, interested purely in statistical properties of events. Many forms of science have shaded over to the narrative end of the spectrum (e.g. Goldstein 1995). Psychiatry and neurology, for example, often depend heavily on case studies, which chronicle the particular life history of individual patients. While science, being a heterogeneous set of practices, cannot be absolutely identified with the purely atomistic end of the spectrum, scientific values and practices do cluster towards atomization. Similarly, the humanities are not unanimous in being placed at the purely narrative end, but humanistic projects do tend to have more of the narrative attributes. This means the division of atomization from narrative is meaningful, at least heuristically.

4.2 *Atomization, narrative, and AI*

Atomization is an essential tool for AI. As Brooks (1986b) states,

> There are many possible approaches to building an autonomous intelligent system. As with most engineering problems they all start by decomposing the problem into pieces, solving the subproblems for each piece, and then composing the solutions.

But because atomization is closely linked with mechanicity, its value must be called into question when the goal is building truly intentional beings. As narrative psychology has demonstrated, when humans try to make intentional behavior meaningful, they use a fundamentally different procedure from that of atomization and the scientific method. Rather, humans create meaning by structuring their experience according to narrative, in the tradition of the

humanities. This difference between the atomistic standpoint of the agent designer and the narrative viewpoint of the eventual agent audience can undermine the designer's ability to construct intentionally understandable agents.

To understand how this works, consider the situation where a designer is building an autonomous agent on scientific principles. S/he starts with a living agent — or idea of an agent — that the designer wants to copy. The designer tries to understand the dynamics of this agent's behavior by finding out its atomic attributes. For example, the designer may try to find out the typical activities in which the agent engages, the conditions under which each activity is likely to occur, and the length of time the agent tends to spend on various activities. Each of these attributes reflects one measurable, objectively verifiable property of the dynamic which the agent designer would like to reproduce.

The designer may objectively and accurately measure many different attributes of the agent. Using these facts, the designer can then construct a system that has the same attributes; for example, the new agent will have the same typical activities, the same probability of being in any particular behavior, and spend the same amount of time on each activity. Once the system can generate behavior that closely approximates the finite list of atomic attributes with which the designer has measured the agent, the designer is satisfied that the agent is a reasonable facsimile of the living agent. Scientifically speaking, the designer is correct.

But now consider the user's point of view. Rather than being interested in the empirically determinable individual attributes of the creature, the user focuses on how the creature's activities seem to meld together into a whole. Rather than caring about whether the creature engages in a particular set of typical activities, for instance, the user may want to know how the activities relate to each other. Rather than caring about the percentage of time spent in any particular activity, the user may wonder why the agent seems to prefer some activities to others at different moments. Rather than caring about the objectively measurable attributes of the original agent, the user will probably care about whether the subjective part of the original agent — its soul or personality — seems to come across.

These *narrative* attributes of the agent's activities — the extent to which the agent's behavior is not simply the sum of predictable parts — are precisely what the scientific copy of the creature has left out. This means that *even if the designer succeeds in making an accurate copy according to scientifically*

measurable properties, from the point of view of the user the living creature is fundamentally different from the constructed agent.

If we are to build agents that truly appear intentional, then, we need to include narrative properties in our design of artificial creatures. Currently, many (though by no means all) AI techniques fall on the scientific end of the spectrum in Table 1. This atomistic worldview reflects itself not only in the internal code of the agents, but also in the quality of the externally observable behavior that forms the basis by which audiences try to understand the agent. The challenge for an AI that wants to build, not just intelligent tools, but intentional agents, is to find ways of moving AI methodology towards the values embodied in narrative. The point is not that narrative is good and science as embodied in current AI is bad, but that we need *both* narrative and AI techniques set in relationship to one another. With the development of a dialectical relationship between narrative and AI, we may finally be able to reach our goal: the development of artificial beings we can truly call intentional.

Notes

1. If the goal of behavior that is not just "life-like" but alive seems trivial or useless to you, you may take heart in the following side-effect of this goal: if we know how humans try to understand intentional behavior, we may be able to build creatures that are more understandable.

2. This is unfortunate, since the original idea of constantly redeciding behavior came in work explicitly interested in diachronic structure. Philip Agre and David Chapman (1987) focus, not on the design of the agent per se, but on the ongoing dynamics of the agent and the environment. The goal is to construct action-selection so that, when put in a particular environment, agents *will* tend to have particular diachronic structure in their behavior. Continuous redecision is part of this work because it keeps the agent's actions closely tied to the agent's context, a property that is also important for narrative, as we will see below. However, the concept of the action-selection algorithm itself tends to undermine diachronic structure, especially when it is used to simply design the agent's behavior alone, rather than for designing a complex interaction between an agent and a dynamically changing environment.

3. Similar arguments may hold for robots. The Sony robotic dogs at Agents '97 were a compelling demonstration that robots may have much greater affective power than even graphically represented agents (Fujita and Kageyama 1997).

4. It may be that one day we can use machine learning to develop this structure instead; whether this learned agent must also be structured abstractly remains to be seen.

5. A similar perspective is used to automatically generate ideologically-based understanding of news stories in (Carbonell 1979). For a humanist example of the effect of generic ways of thinking on the actions we take in our everyday lives, see (Sontag 1979).

6. This is inspired by, but not the same as, the notion of schizophrenia used in psychiatry.

References

Agre, Philip E. 1988. *The Dynamic Structure of Everyday Life*. PhD thesis, Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, MA.

Agre, Philip E. 1997. *Computation and Human Experience*. Cambridge, UK: Cambridge University Press.

Agre, Philip E. & David Chapman. 1987. Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, 268–272.

Analytix, Inc. 1996. http://www.analytix.com/.

Baker, Stephen & Carol Matlack. 1998. Chernobyl: If you can make it here... *Business Week*, March 30, 1998, 168–170.

Bates, Joseph, A. Bryan Loyall, & W. Scott Reilly. 1992. *Integrating reactivity, goals, and emotion in a broad agent.* Technical Report CMU-CS-92–142, Carnegie Mellon University. Also appeared in the *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Bloomington, Indiana.

Blair, David & Tom Meyer. 1997. Tools for an interactive virtual cinema. In Robert Trappl and Paolo Petta (Eds.), *Creating Personalities for Synthetic Actors: Towards autonomous personality agents*, number 1195 in Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag, 83–91.

Bledsoe, Woody. 1986. I had a dream: AAAI presidential address. *AI Magazine*, 7(1):57–61.

Blumberg, Bruce. 1994. Action-selection in Hamsterdam: Lessons from ethology. In *Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior*, Brighton.

Blumberg, Bruce. 1996. *Old Tricks, New Dogs: Ethology and interactive creatures*. PhD thesis, MIT Media Lab, Cambridge, MA.

Brooks, Rodney. 1986a. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2:14–23.

Brooks, Rodney. 1986b. *Achieving Artificial Intelligence through building robots*. Memo 899, MIT AI Lab.

Brooks, Rodney. 1997. From earwigs to humans. *Robotics and Autonomous Systems*, 20(2–4): 291–304.

Bruner, Jerome. 1986. *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Bruner, Jerome. 1990. *Acts of Meaning*. Cambridge, MA: Harvard University Press.

Bruner, Jerome. 1991. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21.

Carbonell, Jaime. 1979. *Subjective Understanding: Computer models of belief systems*. PhD thesis, Yale University Department of Computer Science, New Haven, CT.

Dautenhahn, Kerstin. 1997. I could be you – the phenomenological dimension of social understanding. *Cybernetics and Systems Journal*, 28(5): 417–453.

Dautenhahn, Kerstin & Chrystopher Nehaniv. 1998. Artificial life and natural stories. In *International Symposium on Artificial Life and Robotics (AROB III)*, volume 2, Beppu, Oita, Japan, 435–439.

Dennett, Daniel. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Don, Abbe. 1990. Narrative and the interface. In Brenda Laurel (ed.), *The Art of Human-Computer Interface Design*, Reading, MA: Addison-Wesley, 383–391.

Elliott, Clark, Jacek Brzezinski, Sanjay Sheth, & Robert Salvatoriello. 1998. Story-morphing in the affective reasoning paradigm: Generating stories semi-automatically for use with 'emotionally intelligent' multimedia agents. In Katia P. Sycara and Michael Wooldridge (Eds.), *Proceedings of the Second International Conference on Autonomous Agents*, New York: ACM Press.

Frank, Adam, Andrew Stern, & Ben Resner. 1997. Socially intelligent virtual petz. In Kerstin Dautenhahn (ed.), *Proceedings of the 1997 AAAI Fall Symposium on Socially Intelligent Agents*, Menlo Park, CA: AAAI Press. AAAI Technical Report FS-97–02.

Fujita, Masahiro & Koji Kageyama. 1997. An open architecture for robot entertainment. In W. Lewis Johnson (Ed.), *Proceedings of the First International Conference on Autonomous Agents*, New York: ACM Press, 435–442.

Galyean, Tinsley. 1995. *Narrative Guidance of Interactivity*. PhD thesis, MIT Media Lab.

Goldstein, Kurt. 1995. *The Organism: A holistic approach to biology derived from pathological data in man*. Zone Books, NY.

Hsu, Feng-hsiung, Thomas S. Anantharaman, Murray S. Campbell, & Andreas Nowatzyk. 1990. Deep thought. In T. Anthony Marsland and Jonathan Schaeffer, (Eds.), *Computers, Chess, and Cognition*. New York: Springer Verlag, 55–78.

James, Ellen. 1998. *Her Protector*. Harlequin Superromance. Toronto: Harlequin Books.

Laurel, Brenda. 1986. Interface as mimesis. In *User-Centered System Design*, Hillsdale, NJ: Lawrence Erlbaum Associates, 67–85.

Laurel, Brenda. 1991. *Computers as Theatre*. Reading, MA: Addison-Wesley.

Loyall, A. Bryan. 1997. *Believable Agents: Building interactive personalities*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA. CMU-CS-97–123.

Loyall, A. Bryan & Joseph Bates. 1993. Real-time control of animated broad agents. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science* Society, Boulder, Colorado.

Maes, Pattie. 1989a. The dynamics of action-selection. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, volume 2, pp. 991–997, Detroit, MI:Morgan Kauffman.

Maes, Pattie. 1989b. *How to do the right thing*. AI Memo 1180, MIT AI Laboratory.

Maes, Pattie. 1991. Behavior-based Artificial Intelligence. In Jean-Arcady Meyer and Stewart W. Wilson, (Eds.), *From Animals to Animats 2*, Cambridge, MA: MIT Press, 2–10.

Massumi, Brian. 1992. *A User's Guide to Capitalism and Schizophrenia: Deviations from Deleuze and Guattari*. Cambridge, MA: MIT Press.

Miedaner, Terrel. 1981. The soul of the Mark III Beast. In Douglas R. Hofstadter and Daniel C. Dennett, (eds), *The Mind's I: Reflections on self and soul*. New York: Basic

Books, Inc., 109–115.
Minsky, Marvin. 1988. *The Society of Mind*. New York: Simon and Schuster.
Neal Reilly, Scott. 1996. *Believable Social and Emotional Agents*. PhD thesis, Carnegie Mellon University. CMU-CS-96–138.
O'Neill Brown, Patricia. 1997. Setting the stage for the culturally adaptive agent. In Kerstin Dautenhahn (ed.), *Proceedings of the 1997 AAAI Fall Symposium on Socially Intelligent Agents*, Menlo Park, CA: AAAI Press. AAAI Technical Report FS-97–02. 93–97.
Porter, Tom. 1997 Depicting perception, thought, and action in *Toy Story*. In *First International Conference on Autonomous Agents*. Invited Talk.
Schempf, Hagen. 1995. *BOA: Asbestos pipe-insulation removal robot system*. Technical Report CMU-RI-TR-95–15, Carnegie Mellon University Robotics Institute, Pittsburgh, PA.
Sengers, Phoebe. 1994. Using culture to understand agents. In *AAAI Believable Agents Workshop* (Palo Alto, CA).
Sengers, Phoebe. 1996a. Socially situated AI: What it is and why it matters. In H. Kitano (ed.), *Proceedings of the AAAI-96 Workshop on AI / A-Life and Entertainment* Menlo Park, CA: AAAI Press. AAAI Technical Report WS-96–03.
Sengers, Phoebe. 1996b. Symptom management for schizophrenic agents. In *AAAI-96*. Menlo Park, CA: AAAI Press, vol. 2, 1369.
Sengers, Phoebe. 1998a. *Anti-Boxology: Agent Design in Cultural Context*. PhD thesis, Carnegie Mellon University Department of Computer Science and Program in Literary and Cultural Theory, Pittsburgh, PA. CMU-CS-TR–151.
Sengers, Phoebe. 1998b. Do the thing right: An architecture for action-expression. In *Proceedings of the Second International Conference on Autonomous Agents*, ACM Press, 24–31.
Shakes, Jonathan, Marc Langheinrich, & Oren Etzioni. 1997. Dynamic reference sifting: A case study in the homepage domain. In *Proceedings of the Sixth International World Wide Web Conference*, 189–200.
Sontag, Susan. 1979. *Illness as Metaphor*. New York, NY: Vintage Books.
Steels, Luc. (1994). The Artificial Life roots of Artificial Intelligence. *Artificial Life*, 1(1–2):75–110.
Stone, Brian A. 1996. Dynamically sequencing an animated pedagogical agent. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, OR, 424–431.
Thomas, Frank & Ollie Johnston. 1981. *Disney Animation: The illusion of life*. New York: Abbeville Press.
Various. 1985. Martin Luther (trans.), *Die Bibel*. Deutsche Bibelgesellschaft, Stuttgart.
Vere, Steven & Timothy Bickmore. 1990. A basic agent. *Computational Intelligence*, 6:41–60.
Wise, J. MacGregor. 1998. Intelligent agency. *Cultural Studies*. 12:3.

CHAPTER TWO

Digital Augmentation of Keepsake Objects

A Place for Interaction of Memory, Story, and Self

Jennifer Williamson Glos

1. Introduction

Picture an old teddy bear sitting on someone's shelf. It is missing an ear, most of its stuffing, and looks more like a rag than a stuffed animal. But it will never be exchanged for a new one, nor thrown out, because of its great sentimental value. It has personal and emotional significance, symbolizing someone's childhood and who they once were. I refer to such objects — stuffed animals, photographs, souvenirs, heirlooms — as keepsake objects. They are unique for their ability to hold memories and narratives from the past. Now imagine if you could pick up that old teddy bear and it could tell back to you all the stories you told it when you were growing up — stories from when you were six, eight, twelve years old. What if it could tell you stories of your older sister, mother, grandmother?

An object has two components, its utilitarian purpose, and its underlying emotional and psychological value. Much research at the MIT Media Laboratory and other research institutions is focused on the digital augmentation of objects: moving computation outside of the traditional computer and into the objects of the everyday world. As we make our environment increasingly digital, we have traditionally only considered the practical component of objects. This chapter examines how to enhance both parts of an object: function and personal meaning. In doing so, we will look at the humanist side

of technology, and the psycho-social impact of our inventions.

> The paradox I want to point out is that these objects which always have, in principle, a function, a utility, a purpose, we believe we experience as pure instruments, whereas in reality they carry other things, they are also something else: they function as the vehicle of meaning; in other words, the object effectively serves some purpose, but it also serves to communicate information; we might sum it up by saying that there is always a meaning which overflows the object's use. (Barthes, as cited by Gonzalez 1996, p.70)

Rosebud, the name of the project I will be discussing, comes from the 1941 Orson Welles' film *Citizen Kane*. The film, a biography of the newspaper magnate, William Randolph Hearst, centers on the mystery of Kane's dying word, "Rosebud." Rosebud is revealed to be a sled from his youth, symbolizing the narratives and memories of his lost childhood. The final scene in *Citizen Kane* is a camera pan of a huge room overflowing with all the material objects Kane acquired during his lifetime. Lost among those objects, small and hidden, is the Rosebud sled, of more worth to Kane than all the expensive objects put together, because it was the key to his lost childhood. The loss of the sled symbolized his inability to regain the happiness of his childhood. I am digitally augmenting the Rosebuds of the everyday world.

I begin with the theoretical framework for this project: "tangible media:" the use of everyday objects as computer interface devices, and "keepsake objects:" objects with personal significance. By showing how both the physical and digital worlds are used to support self-identity, I propose digitally enhancing keepsake objects to create a new place for the interaction of memory, story, and self at the confluence of the physical and digital worlds. The second half of this chapter focuses on an implementation of this idea, the Rosebud project. Rosebud uses stuffed animals as the keepsake objects. I describe key attributes of stuffed animals: anthropomorphic, story-evoking, and gender-neutral. These characteristics shaped the design of the system. I conclude with the technical details of Rosebud and user study results.

This chapter is a condensed version of my Master's Thesis (Glos 1997a), completed September 1997, done at the MIT Media Lab in the Gesture and Narrative Language Research Group under Professor Justine Cassell.

2. Theoretical Framework

2.1 *Merging physical and digital worlds*

Rosebud builds upon the design concept of tangible media (Ishii and Ullmer 1997), which includes the use of physical objects as part of the human-computer interaction. In Ullmer and Ishii (1997), a comparison is laid out between levels of abstraction of graphical user interface (GUI) icons and physical user interface icons (phicons). They describe four phicon levels (Figure 1).

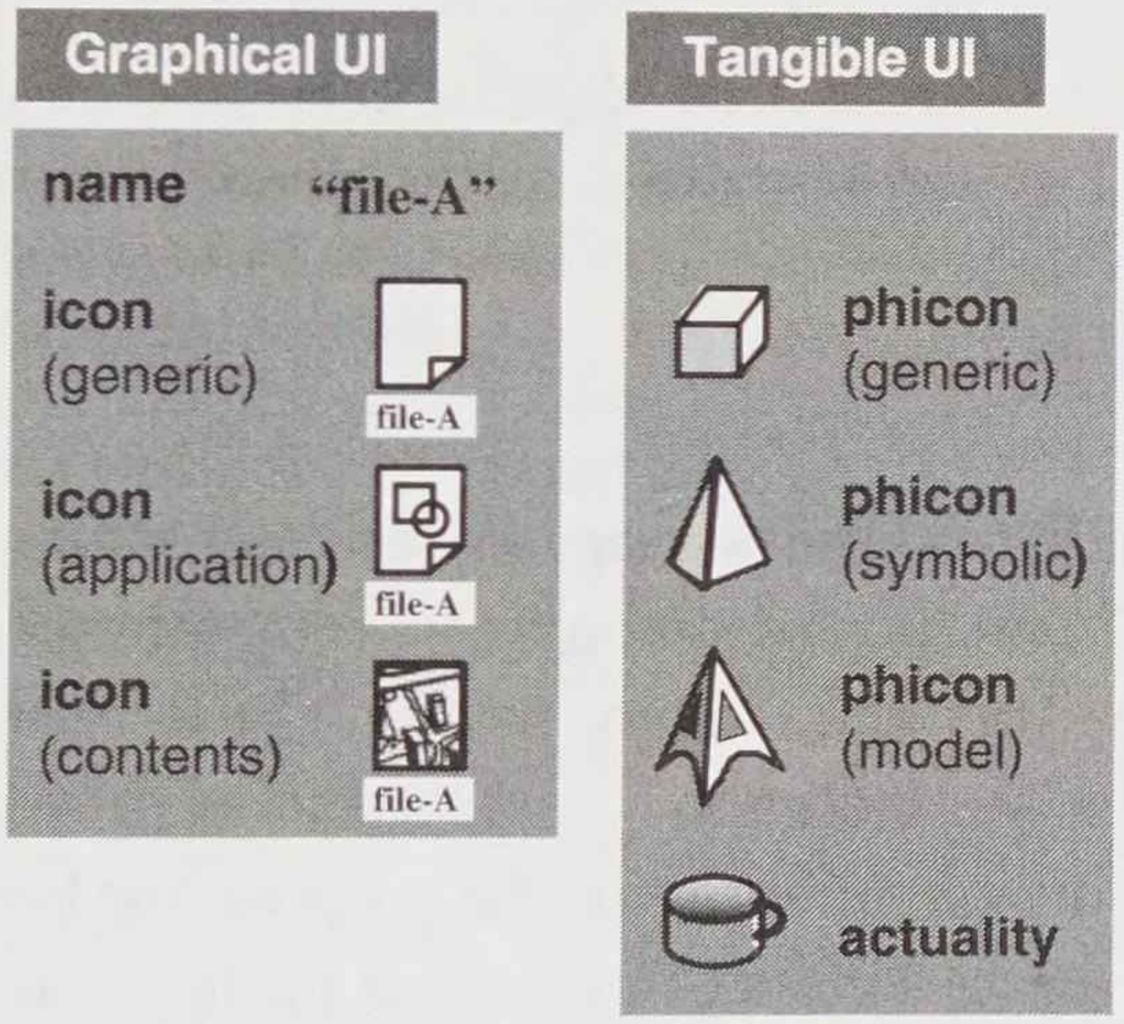


Figure 1. Tangible User Interface Continuum (Ullmer and Ishii, 1997)

(1) generic phicons (e.g. a box) which could represent any data; (2) symbolic phicons (e.g. a pyramid) which represent a class of data, such as geographic landmarks; (3) model phicons (e.g. a mini Eiffel Tower) which represent a specific data instance; lastly, (4) actualities (e.g. the actual Eiffel Tower in Paris), which represent themselves. Actualities do not have a graphical equivalent, because everything in the digital world is an image that represents. As Ullmer and Ishii point out, the Eiffel Tower is a highly inconvenient interface device; however, keepsake objects, by their size and nature, will prove to be ideal for such an interaction. This chapter will illustrate the use of such actualities as interface devices.

Actualities have several characteristics inherent in their nature which make them excellent as interface devices:

Meaningful container: (Houde 1993) introduces the concept of "meaningful containers," where the container reflects the content. "The forms and containers of items that we interact with in the real world provide richer recognition clues than those on our computer desktops."

Physicality: The appeal to the haptic sense allows for an intuitive, natural, and comforting interaction (Druin 1987), permitting interactions such as holding, which are emotionally highly valuable. Secondly, their permanence allows maintenance of form through time, while simultaneously reflecting, in nicks and scratches, marks of usage and passage of time.

The Rosebud project furthers the design concept of "tangible media," through application to a specific sub-set of objects, keepsake objects, and by demonstrating the potential use of actualities as interface devices.

2.2 *Keepsake objects*

Merriam-Webster (1996) defines *keepsake* as "something kept or given to be kept as a memento," and lists as synonyms: "remembrance, memento, memorial, relic, reminder, souvenir, token, trophy." It is formed from two parts: "keep" as in to "preserve, maintain" , and "sake" as in "end, purpose,"[1] with a combined meaning of something with the purpose for keeping. I have appropriated this term to create a class of items that I refer to as "keepsake objects." By my definition, they are a subset of what psychology refers to as *cherished possessions*: "objects belonging to an individual which are considered 'special'." (Dyl and Wapner 1996) Cherished possessions can include electronic equipment, cars, pets, etc. They are one component of a field of psychology which studies objects as mediators between the self and the environment, starting with Winnicott's *transitional object* (1971), a child's first "not-me" object. Keepsake objects are distinct from the broader class of cherished possessions in that they are objects kept over time ("keepsake" versus "cherished"), and the value inherent to them is rooted in their irreplaceability and the personal history associated with it. A keepsake object, if lost, cannot be replaced. Certain objects have a greater tendency to become keepsakes (e.g. heirlooms, souvenirs), but almost any object (e.g., a lace handkerchief) can become one. In contrast, a cherished object may be valued for its monetary value or the function it serves (e.g., clothing), and if replaced with identical items, would be just as cherished.

2.2.1 *Identity in the physical world*

We actively surround ourselves with objects, which aid in the establishing of identity. This grouping of objects can be described as an autotopography, "an array of physical signs in a spatial representation of identity." (Gonzalez 1995: 133)

We symbolically place parts of ourselves in objects. They are a physical embodiment of something otherwise abstract, and are able to serve as *lieux de memoire*, or memory space (Csikszentmihalyi and Rochberg-Halton 1981): handles into personal histories and stories. They counteract the constantly changing world, revealing "the continuity of the self through time, by providing foci of involvement in the present, mementos and souvenirs of the past, and signposts to future goals (Csikszentmihalyi 1993: 23).

> I define keepsake objects as representative of self. Each object plays two roles: the narrative of the object with respect to the self, and the narrative of the object with respect to others (Dyl and Wapner 1996). In the first case, the object serves a reflective role, helping to define oneself to one's self, and anchoring that identity. In the second case, the object serves a presentation function, painting an image of how one wishes to be seen (Goffman 1959).

Csikszentmihalyi (1981) emphasizes the importance of material objects as much more than merely tools. Humans, he says, through creation and interaction with the material world, are able to provide order and a framework of experience to their sense of self and, and by doing so, establish identity.

"Dislocations of rapid historical change, the mass media revolution, and the threat of human extinction," breakdown of communities and authority, and decentralization of our world, have lead us to the *protean self:* a "fluid and many-sided" person, "appropriate to the restlessness and flux of our time." (Lifton 1993) This is a powerful construction of self, but is subject to the danger of no "sustainable inner form," and fragmentation (Lifton 1993, Kellner). Keepsake objects serve an essential role in anchoring the protean self. When belief in ultimate truths is in question and absolutes are disappearing, there is an increased need for continuity. Augmented objects, such as Rosebud, can serve as focusing devices, allowing a person to reflect on who they are over time.

2.2.2 *Identity in the digital world*

As illustrated, humans use the physical world to establish and maintain identity. Interestingly, the digital world is similarly used. (Donath 1996), in

exploring the role of deception of identity in Usenet newsgroups (structured bulletin boards on the internet), confronts similarities and differences between the physical and digital world with respect to presentation of self. She borrows from ethology, presenting two types of identity signals: (1) assessment signals, which are directly manifested from a trait (e.g., large muscles on a human are signals for strength), and are therefore expensive yet reliable, and (2) conventional signals, which are correlated with a trait, but there's no guarantee that the sender has the trait (e.g., wearing a "Gold's Gym Powerlifter" t-shirt), so are less expensive and less reliable. The digital world, likewise, has both types of signals, from domain names in email addresses to homepages. "From the header to the signature, identity cues are scattered throughout..." (Donath 1996: 18) However, "...in the disembodied world of electronic communication identity is problematic, floating free of the stable anchor that the body provides in the real world." (Donath 1996: 1)

While Donath explores digital *expression* of self, (Turkle 1984) asks how computer technology has affected *establishment* of a sense of self. She describes the computer as an "evocative object," showing how children use machines to change their classifications of living versus non-living things, as well as creating a unique definition for human beings. The computer challenges the definition of what it is to be human, not only for Turkle's young subjects, but for adults as well. In the age of IBM's Deep Blue,[2] many ask what the characteristics are which make humans unique.

Donath and Turkle are exploring how the digital world is changing our conceptions of identity. In a series of personal interviews I conducted in spring of 1997, I talked to MIT graduate students about how they establish identity, in both the physical and digital worlds. Their answers illustrate integration of both worlds into their conceptions of self.

Brian and Michael,[3] two of the students interviewed, both have a sense of the importance of the physical objects in their world, discussing the importance of, respectively, their quilt and undergraduate class ring, but they are children of the digital age. They no longer only live in the physical world and no longer do they only have physical bodies. Brian does not center his sense of self in his home or his school; rather than a physical landscape of memories which anchor his identity, he has a digital landscape of memories:

> Everything I had was on the net. I could get to all my old papers. I could get to things I wrote in 7th grade. [During a six year period], almost none of that time was I within a ten-mile or even hundred-mile radius of that machine, but that was sort of where I lived, in this computer in South Carolina.

Similarly, Michael dissociates himself from the physical world and dependence of physical objects. He denies the attitude of collecting, but in the same breath, describes how he collects digitally.

> This whole attitude of collection, of collecting things that mean something. Almost doesn't work for me, because I know in some sense they're just physical things. What's more important to me are the memories associated with them. For example, in the virtual world, I've been collecting all my email for years. That's really exciting to me because fifty years from now, when I look back, I'll actually be able to see what were my thoughts, what were other people's thoughts.

As we inhabit these two worlds, one physical, one digital, so do we establish and express our sense of identity in both places. The digital world is a new arena of expression, with new metaphors and a new language to learn: "The virtual world's subcultures have developed their own patois, with codes and linguistic patterns that identify affiliated participants." (Donath 1996) It has capabilities that the physical world does not, which can allow new expressions of self. It allows identity presentation divorced from the body, which therefore allows expression independent of gender, race, nationality, and physical characteristics. In consequence, there is room for exploration that the physical world cannot provide. It is a networking of the world, allowing near instantaneous communication and publication with the entire world. It is multimedia, allowing expression through a collage of sounds, images, and video. Its most striking, and dangerous, component is its fluidity, and our digital selves are Protean selves, fluid and of many forms, not confined by space or time. It is the postmodern self, where there's "no individual essence to which one remains true or committed. One's identity is continuously emergent, reformed, and redirected."

> If I can speak directly or by electronic mail to a friend in Paris while sitting in California, if I can witness political and cultural events as they occur across the globe without leaving my home, if a database at a remote location contains my profile and informs government agencies which make decisions that affect my life without any knowledge on my part of these events, if I can shop in my home by using my TV or computer, then where am I and who am I? In these circumstances I cannot consider myself centered in my rational, autonomous subjectivity or bordered by a defined ego, but I am disrupted, subverted and dispersed across social space. (Poster 1990: 15)

The digital world, a place defined by its "refresh rate," is by definition always changing. When using digital technology to enhance identity, there is advan-

tage in appealing to the physical for qualities such as permanence.

I have just outlined the many advantages of the physical world, such as constancy, tangibility, and the adding of traces over time. Yet the physical world also has disadvantages for which the digital world can compensate. Sharing of information and distribution through a network are not possible with solely physical objects. They take up space, cannot always be easily transported, and exist in one original form which can be lost or stolen. Unaugmented keepsakes are dependent upon human memory that is notoriously fickle and unreliable. Traditional keepsake devices can not merge or easily become multimedia. Lastly, one can not merge physical and digital selves.

Computationally augmented keepsakes aims to combine the strengths of both worlds, keeping the tangible, physical artifact, but adding to it the stories and memories which can be accessed and altered digitally. In doing so, we are creating a new device for self-expression, which will encourage personal storytelling.

3. Design Context for the Rosebud System

3.1 *Stuffed animal as keepsake object*

My goal is to define a new place for interaction of memory, story, and self, at the intersection of the physical and digital worlds. The Rosebud system is an implementation of an enhanced keepsake object, the stuffed animal. In order to enhance it, we have to understand the stuffed animal's physical affordances which allow it to serve as a prosthesis for memories and stories, and hence, identity. Therefore, we will first look at the characteristics unique to stuffed animals that will make it powerful as a digitally enhanced keepsake. The design process then focuses on strengthening these characteristics.

First and foremost, the stuffed animal is a child's toy. It is unique as a toy in that:

- It is anthropomorphic nature allows the object to share, in an active role, the social experience.
- It is a story-evoking object, not just for memories for the adults, but fictional narratives as well, serving as a creative prop for children's developing imaginations (Goldstein 1994; Sutton-Smith, 1896).
- It is a gender-neutral toy, appealing to both boys and girls (Asakawa and Rucker 1992).

These features will allow us to create a social, interactive, gender-neutral story-telling system.

Secondly, the stuffed animal is a toy that often is kept by adults as well. Interaction styles with them change over time, and the value of a stuffed animal changes with age (Dittmar 1991). While to a child, it is a friend, someone to tell stories about, with, and to, to the adult, it holds the stories they once told and now retell in a new way, with nostalgia.

Stuffed animals, in general, are probably the most likely object from childhood to be kept and to become a keepsake for the adult. "Soft toys were constant companions, much loved and not easily given up, often even in later years." (Sutton-Smith 1986: 47) It is difficult to give up the friend who was always there for you, even if that friend is made of fluff.

Sutton-Smith (1986), in intensive interviews, questions the role of soft toys in the lives of adults.

> Some report keeping the stuffed animals from each stage of their life, as current collections which serve as reminders of their own emotional biography. In childhood it might be a teddy bear; but in adolescence would be "humongous" stuffed animals that were won for them by boy friends, and later "political" animals given by adult friends. (Sutton-Smith 1986: 47)

Sutton-Smith (1986) quotes one women from his interviews who created an autotopography solely out of her stuffed toys:

> Although I did not really have very many important stuffed animals when I was little, I find that I do have several stuffed toys around my apartment now. They have arrived, this cast of goofy characters, in varying stages of the unfolding of my inner life. It is almost as if I have found a comfortable, safe way to physicalize the inner, symbolic ritual of 'regressing' to an earlier, freer state of being. (Sutton-Smith 1986: 48)

Building upon previous work in computational storytelling systems (Cassell 1995), as well as computer-augmented toys (Druin 1987; Umaschi 1997), we are taking an object, the stuffed animal, which already has meaning for us, and which we already understand how to use, and expanding on its inherent storytelling capabilities through computer technology. Rosebud can encourage written storytelling and use of technology in ways that the original stuffed animal could not.

While other systems have explored in detail the use of media technologies and interactive toys for encouraging writing, unique to Rosebud is the placement of such a system into the context of keepsake objects. The child

naturally uses stuffed animals to evoke stories, and the adult naturally uses the stuffed animal as a memory prompt. By creating stories of their own, children find their own voices, and therefore, their selves. When the children are grown, the telling of those stories will become the memories that will form the foundation of identity.

3.2 *Storytelling*

Storytelling does two things: it teaches storytellers about themselves and about the socio-cultural context in which they live, and it makes them better communicators through practice and feedback. Rosebud is an interactive story system that encourages both of these forms of learning.

Objects, such as childhood toys, have rich associations with them, and can serve as a starting point for telling stories and for self-exploration. Children begin storytelling orally, and the use of a stuffed animal as a creative prop is common (Goldstein 1994). Yet the stories from this spontaneous and prolific form of storytelling are usually immediately forgotten and the chance to build upon them is lost. Rosebud is intended to help children progress from oral to written storytelling, using the stuffed animal as a link between the two. The solitary nature of written storytelling, in contrast to oral storytelling with external prompting and the stimulus of others, can be discouraging. With Rosebud, the computer is the listener.

Studies of emergent literacy in children indicate that the features which differentiate oral from written storytelling, such as contextualization, fragmentation (uncompleted sentences, interruptions, over-use of conjunctions such as "and" and "and then"), and improper use of pronomials, are challenges in the acquisition and development of written language (Peterson and McCabe 1996; Martlew 1986). Rosebud has a cognitive and educational goal of improving children's storytelling skills by developing a storytelling system which looks at the user's stories for these common mistakes, and through feedback, modeled after parental scaffolding (Martlew 1986), works with the child to fix these mistakes. The design of this computer-aided editing and revision system is primarily the work of (Sakai 1997).

> Stories and narrative, whether personal or fictional, provide meaning and belonging in our lives. They attach us to others and to our own histories by providing a tapestry rich with threads of time, place, character, and even advice on what we might do with our lives. The story fabric offers us images, myths, and metaphors that are morally resonant and contribute both to our

> knowing and our being known. (Witherell and Noddings, as cited by Lewis 1993: 1)

3.3 *Social experience*

What is the power for children of having their stuffed animal friends come to life? To no small degree, it provides a listener to the child, much like an imaginary friend. When the children are hushed or not listened to by their parents and teachers, they find a new outlet in their toys which, in their minds, come to life and provide confirmation of their alternative world. "They listen attentively for hours, when no one else seems to want to." (Publicity release, as cited by Sutton-Smith 1986: 44) In my user studies interviews, I asked children "If you could make your stuffed animal do anything, what would you want it to do?" The universal answer was "talk to me."

The stuffed animal serves particularly well the role of toy for consolation in loneliness — as a friend. They are often the classic transitional object (Winnicott 1971), such as Linus' blanket, and (Sutton-Smith 1986) describes blankets and teddy bears "...as a general response to the loneliness and search of substitutes brought about by the increasing solitariness of children during the past several hundred years." He hypothesizes that the toys which are the best substitute friends are those with the most humanoid characteristics, such as toys with faces, with voices, or which react.

3.4 *Female voice*

Keepsake objects aid in the expression of gender, one important aspect of identity. Using a computationally enhanced artifact to express identity brings up, but can also help to negotiate the tension in the relationship of boys and girls to technology. (Turkle 1988) argues that women have "computer reticence"," a wanting to stay away because the computer becomes a personal and cultural symbol of what a women is not." Women use technology much less than men as a means of self-expression (Benston 1988), and will even deny that the computer can be a medium of self-expression (Kramarae 1988). While there is a history of women appropriating technologies to use them in ways unintended by their male inventors, men have, historically, been the designers of technology, and technology reflects their interests (Kramarae 1988). By encouraging personal appropriation of computer technology, computers can

be seen as a gender-appropriate — in fact a gender-appropriating — expressive medium. One research objective, therefore, is to design a gender-neutral technology, so that, as a keepsake object, it can be used equally by boys and girls to express themselves.

Cassell (1997) argues that storytelling is an ideal playing field to change the gender imbalance in computer technology. She presents the notion of feminist software design in which authority is distributed to the users, giving them control and voice. The socio-political ramifications of such feminist-designed software are clear. A child's initial introduction to digital technology is through computer and video games. Those who enjoy the games become familiar and comfortable with computers and are more likely to pursue related interests. The games, still designed mostly by men, tend to appeal primarily to male play styles. "Toys are the visible tools of early socialization, preparing children for adult roles (Lipman-Blumen 1984)." This can be seen reflected in the low number of women in computer science and engineering fields (Frenkel 1991). Women, more than men, emphasize the symbolic value of cherished objects for emotional attachment and interpersonal relationships, rather than practical use (Dittmar 1991). Digital keepsakes have the potential to bridge the gender gap by providing an object with both symbolic and practical value. Rosebud, by allowing enhancement of the emotional value of keepsake objects through the computer, presents possibilities for the computer as a tool for girls, designed with respect to the external social environment as well as the cognitive and affective characteristics of girls.

4. System Implementation

4.1 *Overview*

Simply said, Rosebud links children's stories to their toys. It is a three-way interaction between child, computer, and toy. The toy serves to inspire and encourage the child, as well as serving as a memory object, the access point to the stories. The computer is an encouraging listener, as well as a teacher, pushing the child to write, write more, edit, improve.

The child and the computer communicate via the dialogue box on the screen, the dialogue focusing on developing the child's written storytelling skills. The computer and the stuffed animal are presented as familiar with each

other, with the computer recognizing the stuffed animal and having a database of information about it. In return, the computer shares with the stuffed animal the child's stories, so that the animal can respond accordingly. The child-stuffed animal relationship, which is built upon the traditional bond, provides emotional and psychological support. The animal responds to the child's storytelling with lights and motion. While the computer provides a more sophisticated, "adult" encouragement, the stuffed animal encourages the child to write through an affective bond.

As of September 1997, the Rosebud system included six stuffed animals:

- A platypus
- An elephant with lights in its ears
- A Hippopotamus
- A small Winnie-the-Pooh bear
- An Eeyore from the Winnie-the-Pooh stories
- A large Winnie-the-Pooh bear, which can move

4.2 *Designing for multiple voices*

Rosebud is an example of feminist software design, following the principles of feminist pedagogy (Brown 1992) in creating educational software. First and foremost, Rosebud encourages the establishing of voice through an open-ended storytelling framework for the child. It values subjective and experimental knowledge, as well as participatory creativity, supporting writing of personal stories. It promotes collaborative learning, not only among several users and through peer review, but presents the computer as a supportive learner partner, rather than as an authoritative viewpoint, avoiding a hierarchical positioning. Contrary to traditional uses of computers as mathematical machines, Rosebud does not view the child's input as right or wrong, but rather gives only limited structure and guidance. It uses a "language of possibility" rather than a "language of critique (Brown 1992)."

> [Feminist teachers] were not set up as knowers among a group of nonknowers. There were more like part-time assistants. They helped as much as they could by giving information about their experiences. (Finke 1993: 17)

With underprivileged children in mind, the interface design was kept purposefully simple and direct, so that the most novice computer user would have little trouble getting started. The user needs only to know how to read (and even then, minimal reading skills are required) and type (which can be done at any

speed). Since the Story Space is free-form typing, children of all writing levels can tell stories. Overall, I aimed for a design that would lend itself to appropriation by the children. I wanted children with different desires and from different contexts to be able to use Rosebud to suit their needs.

Consideration of female-correlated play styles (Benston 1988; Hascombe 1995; Inkpen 1995a 1995b; Jenkins 1997) was an aspect of the design process, in order to produce a gender-neutral technology. Studies at E-GEMS (Electronic Games for Education in Math and Science) have shown that girls prefer sharing a computer and working collaboratively (Inkpen 1995a). Computers with keyboard and mouse interface are geared towards single-user applications. Rosebud focuses on collaboration by allowing multiple-toy use and multiple-author storybooks, so that several children can write a story together about all of their stuffed animals. Likewise, since the toy serves as a storage device, children can trade their stories by lending their stuffed animals to a friend.

4.3 *Stuffed animal interface*

The stuffed animals contain infrared transmitters and receivers, through Poor's (1996, see Figure 2) iRX board design, as well as limited memory capability so that, when held in front of the computer, the computer can identify the toy, as well as download stories to memory within the toy. The

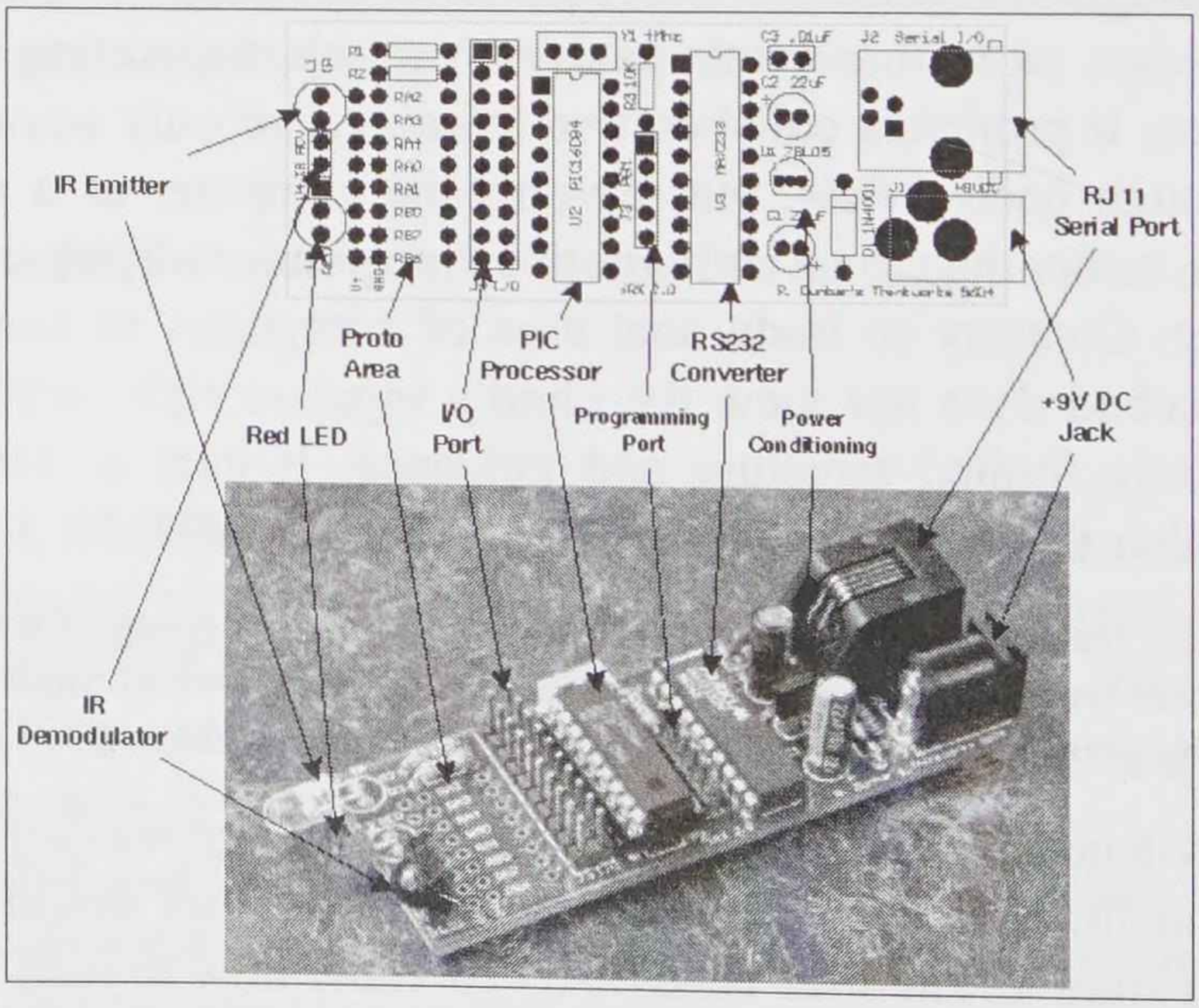


Figure 2. iRX Board[4]

animal is used in two ways: (1) as an index to access stories kept on the computer, and (2) as an actual storage device, like a diskette that reflects its contents.

The characteristics of the stuffed animal which are essential to how children naturally interact with it cannot be lost when augmenting the toy, so the toy as a toy must remain essentially unchanged. All wires and hardware are in the center of the toy's stuffing, safe from harm and unable to be felt, and the toy's plushness and squeezability are unaffected. The toy is not tethered to the computer, and can function independently of it.

The only change that can be seen is the faint, blinking red indicator light from the iRX board to show it is powered and properly working. I placed the board and light in locations where (1) if the animal is held up facing the computer, the machine can easily get a clear signal, and (2) the light could make the most metaphorical sense to the child. The result is that for the elephant, hippopotamus, and both Pooh bears, the light shines from within their mouths. Since the platypus and Eeyore do not have mouths like the others, the lights at the "belly-button" location. I decided to have the light visible for two reasons. First, it simplified the debugging process, immediately notifying me if the battery died or a connection was loose. But, equally important, with every child who used Rosebud, at the outset I would explain how the system works, showing them an iRX board and the red light, and comparing the communication to how a grocery store scanning recognizes items. Without exception, children showed immediate basic understanding and a fascination with the red light. In addition, in several circumstances, when the computer failed to pick up the toy's signal, the children would independently clear away fur, open the mouth wider, or whatever was necessary, to make sure a clean signal could get through. In this manner, no information was "black boxed" or kept hidden from the children.

All six of the Rosebud stuffed animals have this base level functionality: an ability to pick up and transmit signals with the computer, but two prototypes have greater activity and feedback coming from the toy, instead of from the computer. Next are described two features, Story Lights and Motion, which are in progress.

4.3.1 *Story lights*

The elephant stuffed animal reflects how many of the child's stories it knows by way of lights. This has two purposes. First, it provides strong incentive and

motivation for storywriting for the children. Secondly, one important characteristic of keepsake objects is that they are able to reflect wear and tear of usage, while not losing their core nature. The Velveteen Rabbit, for example, developed a bald spot where he was always kissed. With the Rosebud toys, the children's digital interactions with the animals are as important as the physical interactions, however, only the physical is reflected on the toy. In order for the toys to reflect digital usage, in addition to the physical wear and tear, an LED lights up for each story the child tells. To prevent a technical feel to the toy, the lights chosen are very small LEDs which are placed around the outer edge of the elephant ears, giving the impression of "earrings." This completes the metaphor of the animal as an intermediary with the computer, "talking" to the computer through its mouth (the flashing red LED infrared indicator which can be seen there), and "listening" to the child's stories with its ears, which light up with each story. The LEDs are connected to the iRX board inside the toy. Via infrared, the computer tells the iRX board when a story has been added.

4.3.2 *Stuffed animal motion*

The stuffed animal is active only at the start and end of the story-writing process. We are now developing a more interactive stuffed animal. In conjunction with Disney Corporation, and following (Umaschi 1997), we are exploring responsive motion of the toy.

The Pooh Bear stuffed animal moves in several ways: clapping paws, moving ears, nodding head, wiggling nose, or any combination of the above. It receives, via radio signals, frequencies that match to particular motions and the degree of motion. These motions are being linked to the current user activity. For example, when the user is typing a story, Pooh will nod, as though it is registering the story. When users play their recording of the story, Pooh's ears will wiggle as he "listens." Motion will also be linked to particular events, such as when the user types "Pooh" in a story, Pooh will clap, as though happy to hear his name. Motion of the stuffed animal will serve several functions. First, the interaction will more consistently uphold the model of a 3–way interaction between a child, their stuffed animal, and the computer. With the current system, once the story begins to be told, the animal has a passive role, and the only response and feedback is from the computer. The motion will help encourage story writing and editing, but in this case, the encouragement is coming from the toy, rather than the computer. Lastly, the

motion will enhance the fantasy that the toy is "coming to life," giving a more clear indication that the animal is truly "listening" to the child's story.

4.4 *On-screen component*

Rosebud software is written in Symantec C++ for the Macintosh. The on-screen component of Rosebud is comprised of several elements: (1) dialogue space, where the computer talks to the user, (2) story space, where the user types in their stories, and (3) the Storybook, where the stories are archived and can be accessed (Figure 3).

As Glos and Cassell (1997b) and Sakai (1997) have pointed out, Rosebud is well placed for encouraging story writing and revision, and this is what the on-screen component is designed for. Because of the relative novelty of the computer over pencil and paper, as well as the ease and clarity of typing as opposed to painstakingly printing block letters, the computer is an excellent encouraging environment for children's writing. Add to that Rosebud, which provides a traditional children's story-evoking object which communicates

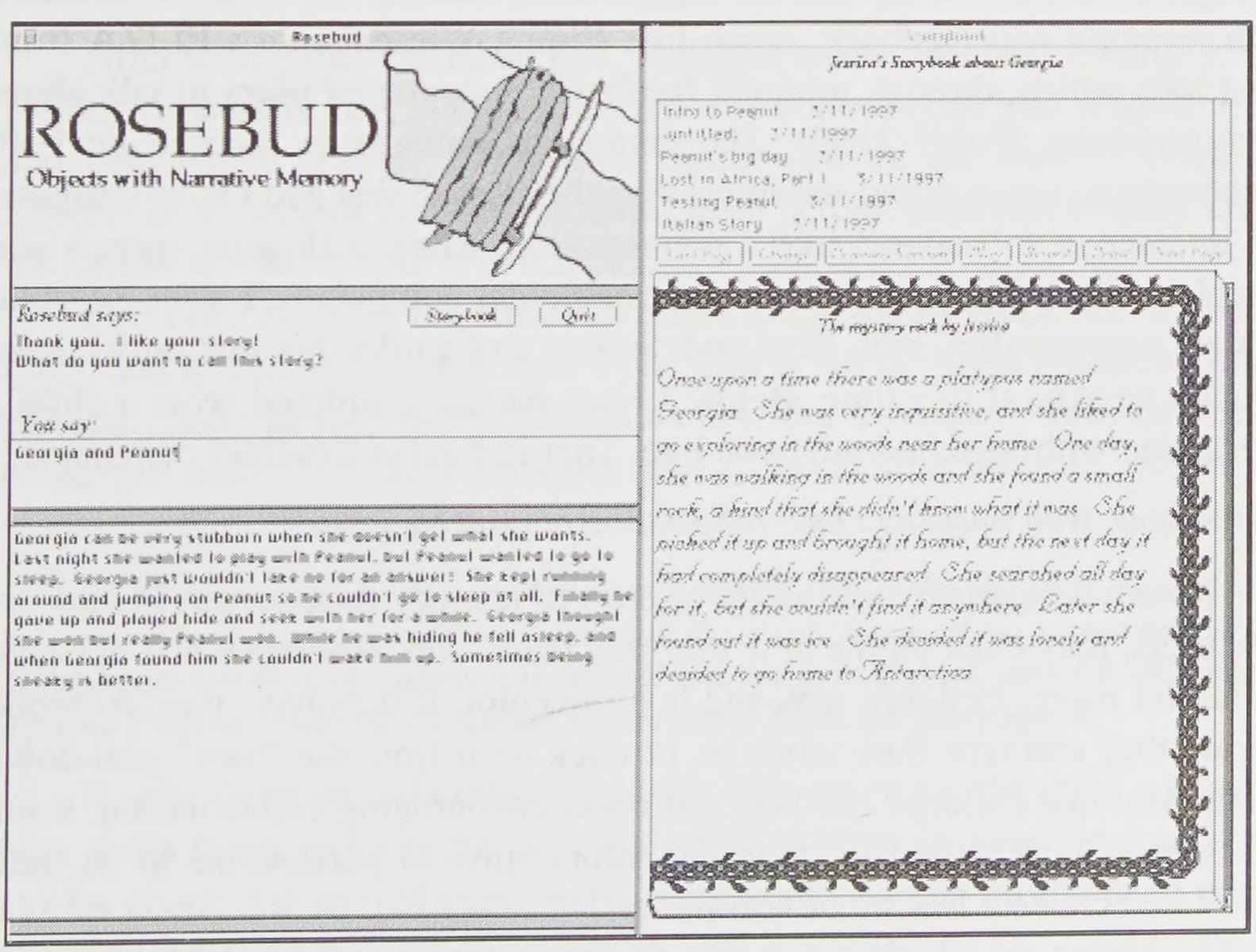


Figure 3. Rosebud On-Screen Interface

and responds to the child. Lastly, motivational features such as the lights on the animal's ears coming on, or the creation of a physical Storybook, urges children to write multiple stories. Once the children are inspired to write, the next challenge is to encourage editing and revision.

Oral enactment of stories is play, but writing of stories is introduced in the classroom environment as a task. Sakai (1997), working with us on the Rosebud project, developed aspects of the Dialogue Space to encourage children's revision through a social, collaborative activity, rather than the solitary environment often found in the classroom. Sakai's story elicitation system is a simple interaction program that branches based on elements present or absent in the user's story, as well as integrates user input into the responses.

4.4.1 *Dialogue space*

Moving from oral to written storytelling can be a difficult task, requiring more abstract thought, and it is a solitary activity, lacking external prompting and encouragement (Martlew 1986). The computer, therefore, guides the user through the story writing and editing process, making it a social interaction with response and feedback, much like Joseph Wiesenbaum's ELIZA from the 1960s which, through minimal feedback, encouraged users to talk about their problems (Foner 1999). The computer dialogue is focused on both encouraging story writing, and improving the story. Sakai and I have designed the interaction to make it like a conversation, using colloquial speech and complete sentences, rather than cryptic computer commands. I want Rosebud to have a personality, with likes and dislikes and quirks, but that only comes through weakly at this time. Below I have partially outlined what a child's interaction with Rosebud might be like. Text in bold is Rosebud comments:

>> Welcome. With whom do I have the pleasure of speaking today?

Rosebud begins by asking the user(s) to identify themselves. If they are new users, they click on the new user box there, and a dialogue box appears, asking for name, birthday, age, and favorite color. If they have used Rosebud before, they can type their name in, or click on it from the "user" pull-down menu. Multiple children can log in at once, encouraging collaboration. From the beginning, Rosebud is gathering information to personalize to an individual and create a feeling of familiarity.

>> Hello, [Jen]. Do you have an animal friend with you today?

Rosebud can be used either with or without stuffed animals, in case the child does not have a toy or, if the children can not agree on which animal to use, they may choose to create their own character.

>> Please show me who you are playing with today. <or> Who are you going to tell a story about today?

Now Rosebud is requesting to "see" the stuffed animal, at which point the child holds the animal up in front of the computer and it reads the infrared signal. This is the heart of Rosebud: the communication between the traditional, familiar, soft and friendly toy and the "box" of the computer which may seem, depending on the child, unfriendly and unfamiliar. By establishing that the toy and the computer talk and are friends, the already accepted and loved toy can pull the computer into the child's acceptance. Druin (1987) approached the issue of the non-child-friendly computer interface by placing the computer *inside* of a giant stuffed animal named Noobie, hiding it away. In this case, rather than hiding it, I am showing children that it can be their friend, just like their stuffed animal is their friend and it is friends with the animal.

>> What is the name of this [type of animal, e.g. hippopotamus]?

After seeing the stuffed animal, Rosebud wants the child to name it. The naming process is key, allowing the child to control (Dyl and Wapner 1996) and personalize (Papert 1996) their possession. In the case of Pooh and Eeyore, or an animal that the child has used before, Rosebud already knows the names and the question is skipped.

From this moment on, Rosebud refers to the toy by name (as well as the user), and it understands that Jane's platypus is not the same animal as John's platypus, and that they have different names. The computer remembering and using the name the child provides gives the child a sense of power and assurance that their decisions are followed.

>> Isn't this your friend, [X, e.g. Plattie or Peanut]? This is the last story I heard about [x]. <or> Why that's my good friend, [Winnie-the-Pooh/ Eeyore]! An author named A.A. Milne wrote lots of stories about him. Would you like to read one? <or> Oh, this must be a friend I haven't met. Please tell me a story about [x] in the Story Space.

Rosebud can provide other stories as inspiration for the child's imagination. With Pooh and Eeyore, it is a story by Milne. If the child has written stories about this animal previously, the computer pulls up the most recent, showing the child that it remembers, as well as reminding the child of the last story they told, in case they would like to build on it. In addition to stories,

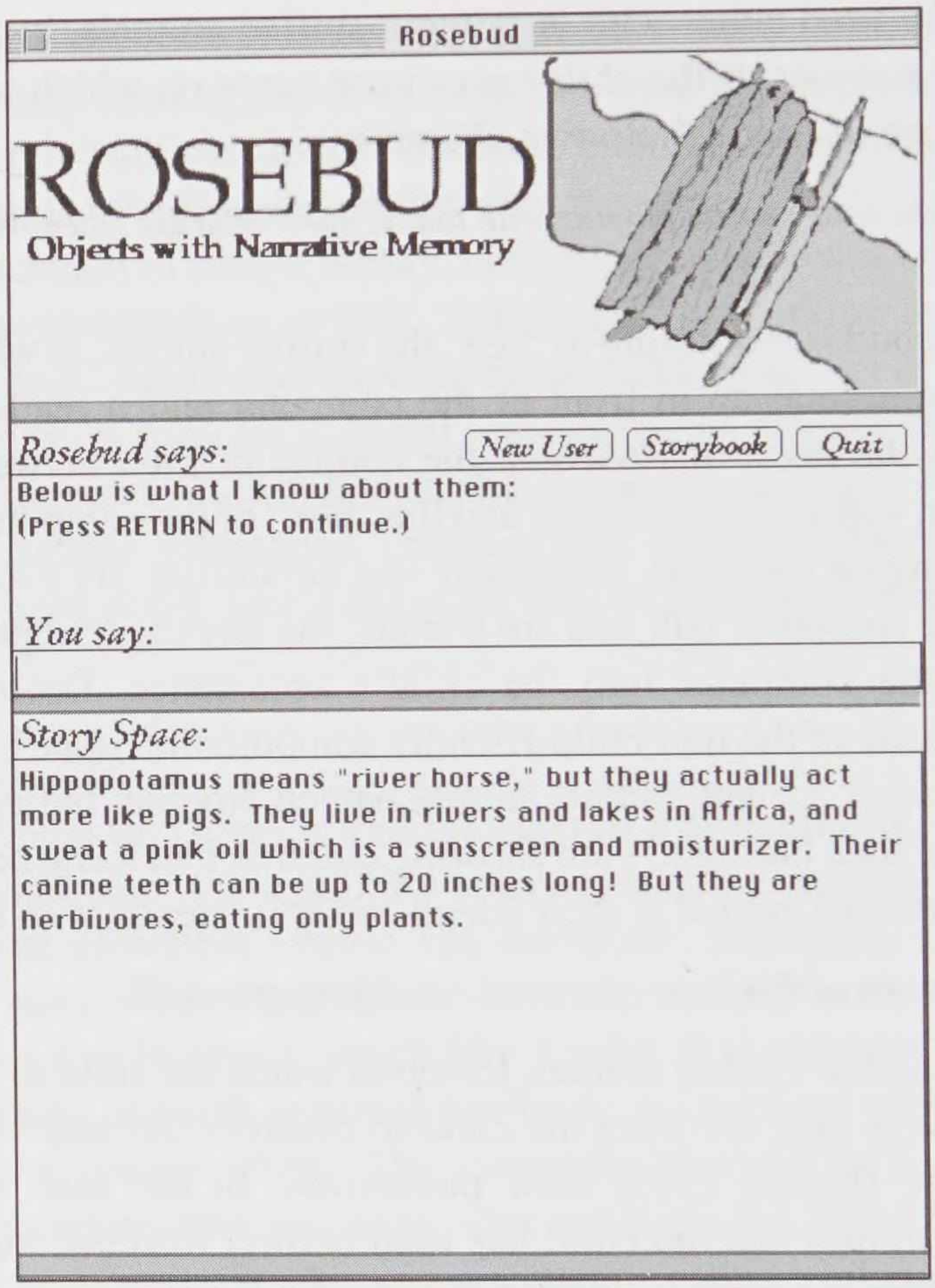


Figure 4. Rosebud Gives Animal Facts in Story Space

Rosebud also has a collection of stories and tidbits of information about all the animals, so it can appear a knowledgeable friend, as well as personalizes the exchange to the particular toy (Figure 4).

The computer remembers if the user has played with other animals in the past, and asks the user if the different animals have been introduced to one another. This builds up a three-way relationship and causes the child to think about the relationships between the individual animals. They have to begin to create a mental model of who the animals are and who they know. The user(s) can play with multiple animals simultaneously. If there is only one user, this encourages the child to explore social dynamics and relationships between the different animals, developing a community of toys. If there are several users, this allows each child to include their own animal, encouraging collaboration,

and giving every child a voice and a part of the story (the part about their animal) that they control.

After the initial conversation between Rosebud and the user, the computer asks for a story. Hopefully, by this point, the child is relatively at ease with Rosebud and will willingly type in a story. But in case they are still shy, the computer can react accordingly. The conversation to date serves as story fodder and helps the child find a starting place. Many users have done just that with Rosebud, writing stories about how the animal got its name, how it meet the other animals, incorporating information they learned about the animal (e.g., that platypuses live in burrows), incorporating elements from Milne's story world, such as some of the other characters.

As the user writes their story, the system gives generic encouraging feedback such as "I like that story." It looks at the length of the story and, if short, says, "Don't be shy! Please tell me more." This is based off of ELIZA's encouragement.

After writing the story, the user is asked for a title, another chance for personalization and control. The title is used to index the stories in the Storybook.

The most repeated request of children in early testing was to hear their stories told back to them by their stuffed animals. The design option was to either have a computerized monotonic reading, or to record the children's reading of their own stories for later playback. We chose to have the children record their stories and then listen to themselves. This decreased the power of the toy as a separate social entity, but increased its power as a physical and virtual holder of stories. While this is merely a computer tape recorder, the recording and playback was placed in the context of the social storytelling process, and therefore encouraged story revision.

Upon finishing a story, children are asked by the computer: "Would you like to read your story to me?" If they agree, a dialog box appears with record, stop, play commands. They are able to record, listen to their recording, and re-record, until they are satisfied. This plays upon the link between oral history and keepsake objects, and upon children's love of repetition and hearing their own voices. Most importantly, it places the child in both the roles of the author and the listener. When the children read their story aloud, they see and hear errors that would otherwise have gone unnoticed, or from having spoken, they know better what they want to say. This is therefore one entry point into the editing dialogue, where the computer asks a variety of editing questions and

the user has the option to go back and make changes.

The computer's function is to guide the child, unfamiliar with revision, to teach them how to read their own work carefully. Sakai (1997) implemented three editing prompts. First, general questions are asked, geared to focus the child to examine particular sections of the story, such as asking if the user is happy with the story's end. Second, the system runs a function to check the beginning of the story (first eight lines) for words and phrases indicating context, such as "today" or "in the." If it fails to find any, it asks for contextualizing information, such as when and where the story occurred. And last, it encourages peer revision by recommending the user ask another to read and comment on the story.

These editing prompts serve to improve children's storytelling skills. External prompting has been illustrated to advance children's narrative skills, but not only the fact of prompting, but the prompt itself, can make a difference. Rosebud addresses common errors made by children learning to write, in particular, lack of contextualization for an absent reader. Modeled after parental scaffolding (Peterson and McCabe 1996), Rosebud elicits such information if the child forgets it, teaching them this storytelling skill: "context-eliciting questions … foster context-setting skills in children." (Peterson and McCabe 1996)

Wording of this interaction is guided by feminist pedagogical principles, avoiding a right/wrong philosophy and authoritative wording, but instead making suggestions or professing its own confusion and asking the user to clarify. Rather than the editing being based off absolute criteria (e.g., correct spelling and grammar), it is based on the child's liking and satisfaction with their story. The style is also modeled off the closed-ended, information-rich questions that showed to encourage storytelling in young children (Peterson and McCabe 1991).

This revision sequence is activated in several scenarios. First, revision prompts are activated after the users read their story aloud, which places them in the role of the listener, and can bring to attention story errors previously unnoticed. Second, it is also beneficial to review a story with "new eyes" after some time has passed. Therefore, delayed revision is encouraged by the system.

4.4.2 *Story space*

The story space is a free-form text input box, empty with a scroll bar to accommodate any length story. Unlike many "story writing" commercial

products available, we have not provided any structure or content, such as "story starters," pre-made characters, or scenarios, for the child's storytelling. Rather, the objective is to give the child complete authoring power. This can be intimidating at first, which is countered with the Rosebud personality and the stuffed animal. The story space was made purposefully large, taking up a great deal of the screen. This emphasizes its importance and provokes a desire to put something into it.

4.4.3 *Storybook*

The stories are archived in the Storybook (Figure 5). The stories are indexed by author(s), toy(s), and time. For example, if Marcy and Roberta are telling a story about Peanut and Plattie, then the "Storybook by Marcy and Roberta about Peanut and Plattie" will appear. Those stories also show up in the individual books of Marcy and Roberta, marked as co-authored.

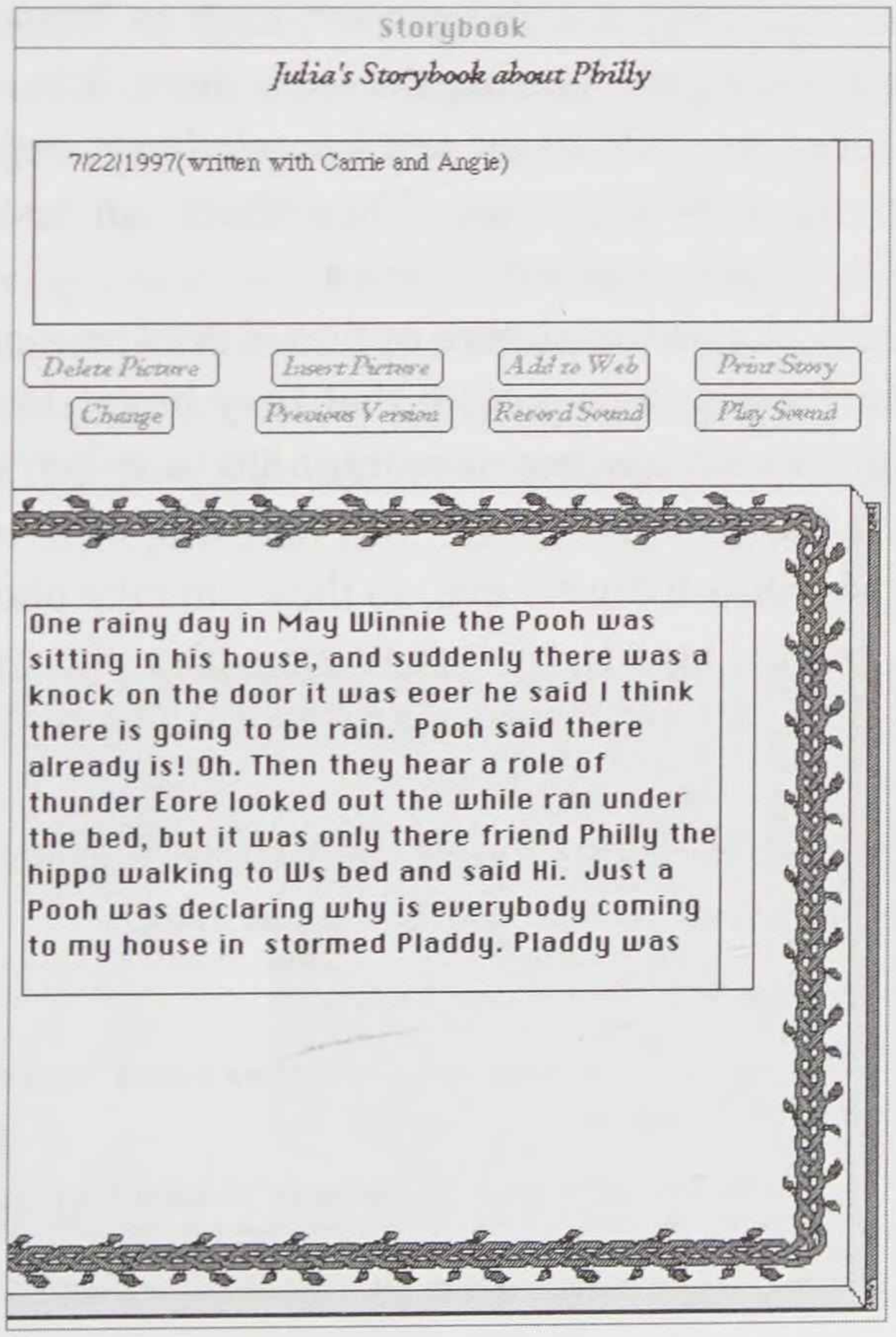


Figure 5. Rosebud Storybook

From the Storybook, the children can do a variety of activities with their stories: (1) add stories to the web, (2) print them out, (3) add a picture, (4) record or play an audio segment to go with it. Also, for editing and revision purposes, whenever users change their story through the "change story" button, the older version of the story is automatically saved. Older versions can be flipped through by clicking on the "previous version" button, allowing the child to see their progress or show changes to another.

Adding to the web and adding pictures provides the next step in complexity with Rosebud. Papert (1996) discusses the importance of extendibility of a project so that the learning potential is never-ending. As loosely-constructed storybook options, each using different software and each with a tangible result, and all adding to the final result, I provide focus for learning new software, as well as motivation for doing a longer-term project.

When the user clicks on the web option, a pop-up menu appears (include image here) which gives the child directions for putting their story on the web page. This requires launching a web browser, such as Netscape, going to the correct address, and cutting and pasting the story there. It is a relatively simple process, but introduces the child to the key knowledge to begin web browsing: the web browser and web addresses. They then can see their own story "published" as well as look at others' stories.

When the user clicks on the picture option, a pop-up menu appears giving directions for adding a photo or picture that they have already created. This function was designed with teacher or mentor aid in mind, who can show the child (1) drawing tools, such as Kid Pix or Photoshop, to helping them create a drawing, (2) the Macintosh digital eye, so they can take photos of themselves and their animals, and (3) how to capture images from the web and save them. Rosebud merely allows the chosen image to be added to their Storybook (Figure 6).

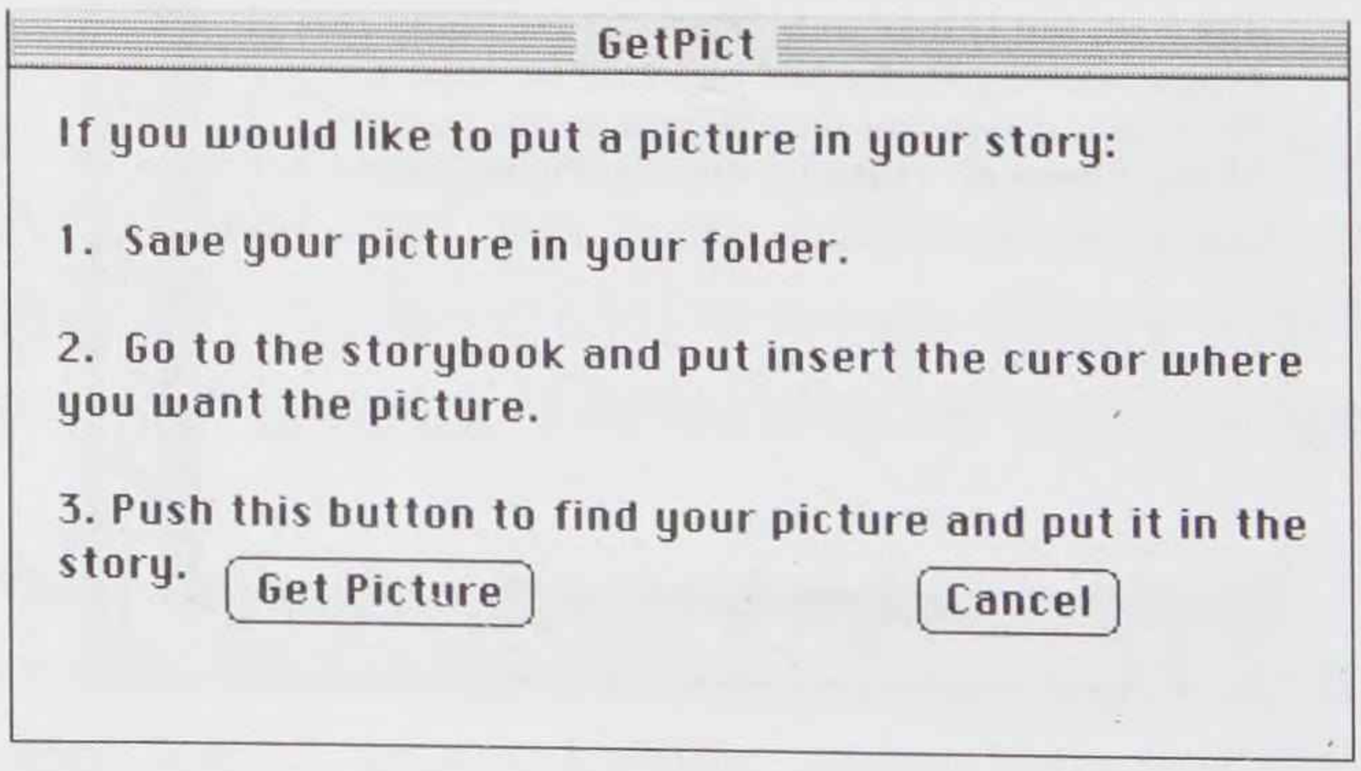


Figure 6. Dialogue Box for Adding Images

Rosebud is designed to give children of all types a voice — to encourage them and convince them that they have something of worth to say. First, they are their own audience when they listen to their own voices. Then, they are encouraged to use peer review and show their Storybook to their friends, who become their second audience. Next, they can print out their stories, showing them to parents and teachers and those who do not have computers. Lastly, they can put their stories on the web, where their audience is the world.

5. User Experiences

During development, approximately fifteen children, ranging from seven to twelve years of age, used Rosebud prototypes, giving early feedback. Overall responses were favorable, indicating a good basic concept. Prompting was successful in encouraging storytelling from initially reticent children, and they were very enthusiastic at the interaction of the computer and toy. Based on their desires, we added the audio capability, and changed the interface to clearly delineate between story writing and user dialogue with the computer. Upon completion of the Rosebud system, more extensive user studies were conducted, involving children over a period of several weeks as part of two after-school programs. Discussing the full results of these studies is beyond the scope of this chapter, however, I will mention a few of the highlights.

Having a stuffed animal as an intermediary between the children and the computer had several positive effects. Familiarity and comfort with the stuffed

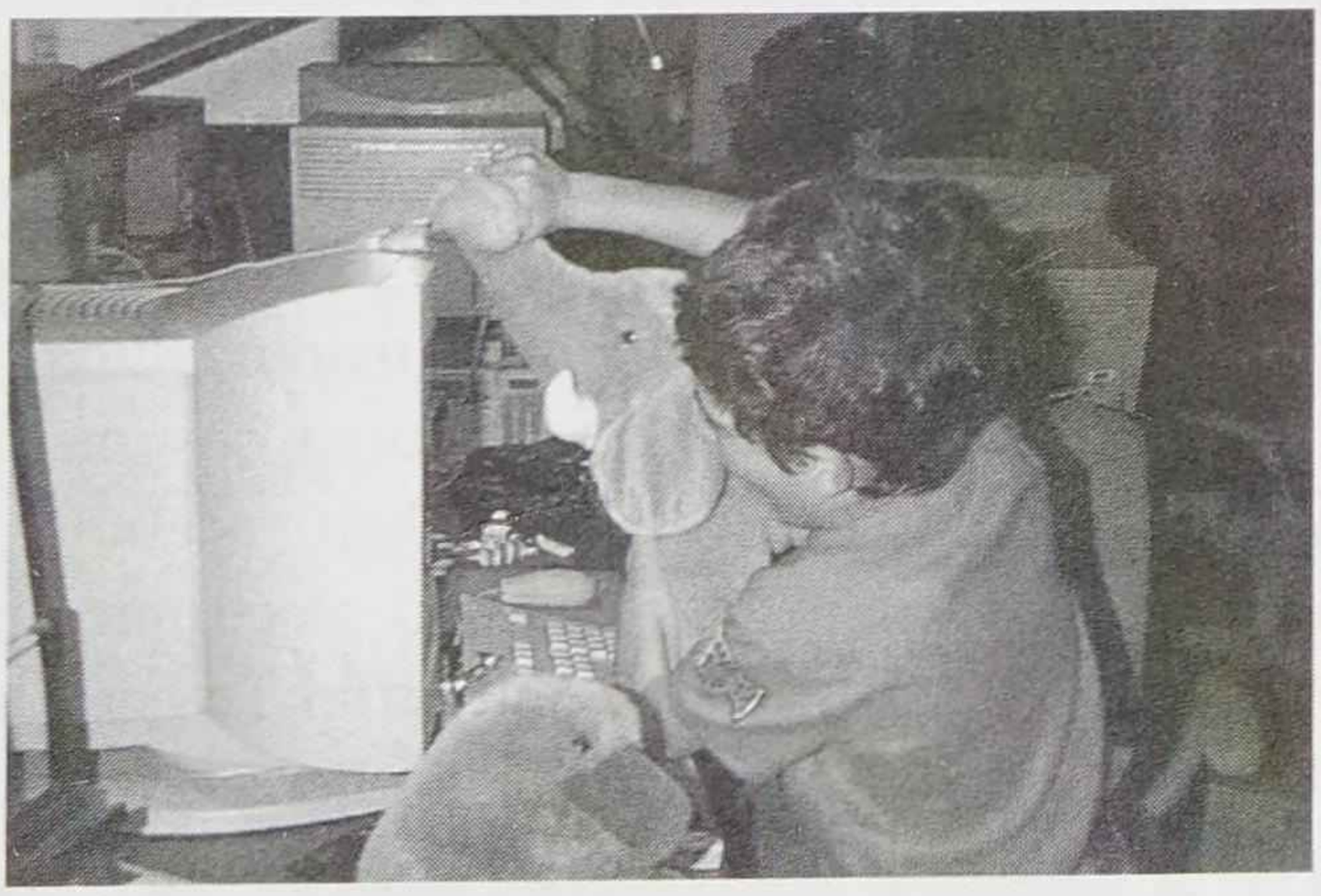


Figure 7. User helping Rosebud elephant to "talk" to computer

animal, as well as ease-of-use, successfully drew in the initially hesitant and technically inexperienced children (Figure 7). In addition, the children's natural appropriation of stuffed animals carried over into their use of the Rosebud system, and it appeared that they began to understand what is perhaps most important about computers: "you can use them for your own purposes." (Papert 1996: 27).

Without doubt, the recording and playback of their stories was one of the children's most enjoyed activities, plus the audio proved to have a powerful side-effect of encouraging story editing, since the children craved perfection in their recordings. Perhaps most significant was the effect Rosebud had on their impression of written storytelling. Whether or not they initially said they liked to tell stories, afterwards they called Rosebud "playing" or "a game" and one child said, "I learned that making stories is fun." Rosebud successfully moved story writing from the realm of work into the domain of play, even the laborious and despised process of editing.

Rosebud allows collaborative work through simultaneous multiple-toy use and multiple-author storybooks. For example, two children can each show the computer a toy at the same time and Rosebud will structure the interaction and storybook to include both animals and both authors. We compared two children working together using one, two, and no animals. Control was most often linked to possession of the keyboard, such as in one case, a child taking it by force from their partner in order to control the story. However, the toys helped to distribute the control. Possession of an animal gave storytelling authority, and when the child without the keyboard had an animal, they were more likely to participate and actively influence the story. Collaboration was strongest with two animals, and weakest with no animals.

Lastly, these studies answered a core paradox of Rosebud: the concept of a manufactured keepsake object. Organic keepsake objects, such as your grandmother's handkerchief, are typically not purposefully designed to become keepsakes. They contrast with the mass manufactured souvenirs to be found in stalls on the street corner of any tourist destination: hundreds of small, plastic Eiffel Towers, to remind one of their vacation.

Rosebud is manufactured, with chips and electronics placed inside to enhance the keepsake nature. But, unlike the Eiffel Towers, an unused Rosebud has no meaning before someone adds his or her stories. It is not meant to generically represent a class of experiences, such as "my vacation in Paris," but is more of an empty book, placed inside a highly organic object likely to

become personally valuable to an individual. Like a camera, it is the tool for creating keepsakes. A camera is meaningless until someone takes a photograph, and that photograph is *theirs*, not manufactured or mass marketed.

In Williams's *The Velveteen Rabbit* (1984: 6), the rabbit asks the skin horse what is it to become real:

> "What is real?" asked the Rabbit one day..."Does it mean having things that buzz inside you and a stick-out handle?"
>
> "Real isn't how you are made," said the Skin Horse. "It is a thing that happens to you. When a child loves you for a long, long time, not just to play with, but REALLY loves you, then you become Real."

Likewise with keepsakes: it is not gears that click and turn — is it not the technology — which is important. It is if children love the Rosebud animals, and take them into their hearts. In these user studies, the children gave life to Rosebud by giving it their stories.

6. Future Work

While in academia projects such as Rosebud, Noobie (Druin 1986), and SAGE (Umaschi 1999; Umaschi 1997) are exploring how stuffed animals and computers augment one another, the toy and computer industries are also merging new technologies with traditional toys. In the fall of 1997, Mattel released *Talk with Me Barbie*, a Barbie doll which can communicate with a child's own mini-computer via an infrared signal in the doll's necklace. It personalizes to the child with name, birthday, favorite color, and when Barbie talks, her lips move. The child's computer can also plug into a regular PC where a CD-ROM can be used to accompany it. Simultaneously, Microsoft released, for two to five year olds, the first of its "ActiMates" characters: a Barney stuffed animal which can function alone or, with appropriate hardware, can communicate with a TV/VCR or home computer via radio signals. Alone, it moves, talks, sings, and responds to touch or changes in lighting. With the videotape, it becomes a "viewing companion," commenting on the show. With a PC, it reacts to several companion CD-ROMs (Microsoft News Release).

These systems excel at the development of interactive toys as collaborative social entities. The design objective of Rosebud is more than another

illustration of an augmented children's toy, but rather the first step in exploring a new type of keepsake object, namely digitally augmented keepsakes, with both physical and virtual components. Future work needs to focus on this aspect of the system, by strengthening the story storage components and studying the effectiveness of Rosebud as a memory device over a period of time. In addition, building a second system using a different keepsake object would allow contrast and comparison.

In addition, a difficulty with using digital technology for archiving purposes, such as storing information with a keepsake object, is that rapid technological advances and inconsistent backwards compatibility makes it likely that the data will be difficult to access in the future. The solution is to make the system self-sufficient and highly stable. In the case of Rosebud, the toy should function independently of the computer, particularly playback. This could be done with logical access to the audio recordings, such as potentially though interaction with the story lights.

7. Conclusions

In this chapter I present the concept of digitally augmented keepsake objects as a merging of the physical and digital worlds to support self-identity. I outline how the physical world, traditionally, has served to aid in self reflection, and how we might take advantage of the elements of the material world which allows this, and build upon them computationally. I then narrow my focus to one instantiation of such a system, Rosebud, which uses stuffed animals, binding children's stories to them.

By digitally augmenting features common to all keepsakes: tangibility, emotional value, and storage of stories and memories, I built a digitally augmented keepsake object. More specifically, by enhancing stuffed animals' inherent characteristics, I developed an interactive storytelling system for children of both genders. With this prototype system, I was able to test and understand the advantages of digital augmentation of keepsake objects.

Rosebud illustrates how the marriage of keepsake objects, valued for their strong emotional power and evocative nature, with the computer, can allow the computer to better satisfy users' social and cognitive needs. Rosebud provides a place for the interaction of story, memory, and self. It aims to aid in the development of children's storytelling skills, gives greater power to an object

as a memory device, and allows for development of one's voice and, therefore, one's self. Its functionality changes over time as a user grows from a child, learning how to write stories, to an adult, wanting the stuffed animal to be a memory object of childhood. This work demonstrates one application of tangible media, where the physical component, a stuffed animal, appeals to girls and is a story-evoking object, while the digital component provides memory and a framework to encourage children to explore storytelling further.

Tangible media, or computationally enhanced environments are the future. While there is research on the formation of identities in the digital world (Donath 1996), and work on the formation of identity in the physical world (Csikzentmihalyi 1981), how will these meet in a computationally enhanced environment? This chapter explores how tangible media will need to look at the affective nature of objects, as well as their functional value. With digital media, unanchored and fluid, meeting the Protean self, there is a need for computationally enhanced keepsake objects as a place of interaction of memory, story, and self: where an individual's history can be stored and serve to stabilize self-identity.

Acknowledgements

Thanks to the Massachusetts Institute of Technology for granting the right to republish portions of my 1997 Masters Thesis in this chapter. This work was carried out under the primary guidance of Justine Cassell, Assistant Professor of Media Arts and Sciences at the M.I.T. Media Lab, as well as Hiroshi Ishii, Associate Professor of Media Arts and Sciences at the M.I.T. Media Lab, and Henry Jenkins, Professor of Literature and Director of Film and Media Studies, M.I.T. I would like to thank Matthew Sakai for technical development, as well as undergraduate research assistants Tara Andrews, Kartik Mani, Joy Nicholson, and Irene Wilson. WDI Cambridge made Pooh Bear's motion possible. Thanks as well to editors Manish Tuteja and Brygg Ullmer. This research was supported in part by the Media Lab *Things That Think* and *Digital Life* consortia, and the Grauso Graduate Fellows program.

Notes

1. Webster uses the illustrative example: "for the sake of argument."
2. The first computer to beat the world's presiding human chess champion.
3. All names have been changed.
4. Figure from Robert Poor's iRX web page: http://ttt.media.mit.edu/people/r

References

Asakawa, G. and L. Rucker. 1992. *The Toy Book.* NY: Alfred Knopf.

Belk, R.W. 1991. The Ineluctable Mysteries of Possessions. To Have Possessions [Special Issue]. *Journal of Social Behavior and Personality*, 6 (6), 17–55.

Benston, M.L. 1988. Women's voices/men's voices: technology as language. In C. Kramarae (Ed.), *Technology and Women's Voices.* New York: Routledge & Kegan Paul.

Brown, J. 1992. Theory or Practice — What Exactly is Feminist Pedagogy? *Journal of General Education,* 41, 50–63.

Cassell, J. 1997. Storytelling as the Nexus of Change in the Relationship between Gender and Technology. Speech, *From Barbie to Mortal Kombat: A Conference on Gender and Computer Games.* Cambridge, MA.

Cassell, J. 1996. Feminist Approaches to Software Design: Building Interactive Story Systems for Girls. Speech, *Small Computers in the Arts Network '96.*

Cassell, J. 1995. *Renga: The Cyberstory*, MIT Media Lab, WWW project. http://renga.www.media.mit.edu/child-1010.html

Csikzentmihalyi, M. and E. Rochberg-Halton. 1981. *The Meaning of Things: Domestic Symbols of the Self.* New York: Cambridge Univ. Press.

Csikszentmihalyi, M. 1993. Why We Need Things. In S. Lubar and W.D. Kingery (Eds.), *History from Things.* Washington, D.C.: Smithsonian Institution Press.

Dewey, J. 1938. *Experience and Education.* New York: Collier Books.

Dittmar, H. 1991. Meanings of Material Possessions as Reflections of Identity. To Have Possessions [Special Issue]. *Journal of Social Behavior and Personality*, 6 (6), 165–186.

Donath, J. 1996. Identity and Deception in the Virtual Community" To appear in P. Kollock and M. Smith (Eds.), *Communities in Cyberspace.* Berkeley: University of California Press.

Druin, A. and C. Solomon. 1996. *Designing Multimedia Environments for Children.* New York: John Wiley & Sons.

Druin, A. 1987. *Building an alternative to the traditional computer terminal.* Cambridge, MA: MIT Master's Thesis, Media Laboratory.

Dyl, J. and S. Wapner. 1996. Age and Gender Differences in the Nature, Meaning, and Function of Cherished Possessions for Children and Adolescents. *Journal of Exper. Child Psychology*, 62, 340=377.

Finke, L. 1993. Knowledge as Bait: Feminism, Voice, and the Pedagogical Unconscious. *College English*, 55, 9–27.

Foner, L. 1999. Are We Having Fun Yet? Using Social Agents in Social Domains. (This volume).

Frenkel, K. 1991. Women and Computing. *Communications of the ACM*, 33 (11), 34–47.

Garvey, C. 1990. *Play.* Cambridge, MA: Harvard Univ. Press.

Glos, J. 1997a. *Digital Augmentation of Keepsake Objects: A Place for Interaction of Memory, Story, and Self.* MIT Master's Thesis, MIT Media Lab, Cambridge, MA.

Glos, J. and J. Cassell. 1997b. Rosebud: A Place for Interaction of Memory, Story, and Self. Aizu, Japan: *Cognitive Technology '97 Conference Proceedings.*

Glos, J. and J. Cassell. 1997c. Rosebud: Technological Toys for Storytelling. Atlanta, Georgia: *CHI '97 Conference Companion*, Short Papers.

Glos, J. and M. Umaschi. 1997d. Once Upon an Object: Computationally-augmented Toys for Storytelling. Gold Coast, Australia: *ICCIMA '97 Conference Proceedings.*

Glos, J. and B. Ullmer. 1996. *POEMS: Physical Objects with Embedded Memories,* unpublished video.

Goffman, I. 1959. *The Presentation of Self in Everyday Life.* NY: Anchor Books.

Goldstein, J. (Ed.). 1994. *Toys, Play, and Child Development.* Cambridge, UK: Cambridge Univ.

Gonzalez, J.A. 1996. *Siting Histories: Material Culture and the Politics of Display in the Work of Fred Wilson, Pepon Osorio, and Amalia Mesa-Bains.* Doctor of Philosophy Thesis. Santa Cruz, CA: University of California, Santa Cruz.

Gonzalez, J.A. 1995. Autotopographies. In G. Brahm Jr. and M. Driscoll (Eds.), *Prosthetic Territories: Politics and Hypertechnologies.* Boulder, CO: Westview Press.

Hascombe, B. Beyond Chrome and Sizzle. 1995. *Game Developer,* February, 1995.

Houde, S., and G. Salomon. 1993. Working Together Towards Rich & Flexible File Representations. In *Proc. of INTERCHI'93*, Adjunct Proc., 9–10.

Inkpen, K., Booth, K., Gribble, S., and M. Klawe. 1995a. Give and Take: Children Collaborating on One Computer. In *CHI'95 Proceedings,* Short Papers.

Inkpen, K., Booth, K., Klawe, M., and R. Upitis. 1995b. Playing Together Beats Playing Apart, Especially for Girls. In *CSCL '95 Proceedings.*

Inkpen, K., Upitis, R., Klawe, M., Lawry, J., Anderson, A., Ndunda, M., Sedighian, K., Leroux, S., and D. Hsu. 'We Have Never-forgetful Flowers in Our Garden:' Girls' Responses to Electronic Games. World Wide Web document.

Ishii, H. and B. Ullmer. 1997. Tangible Bits: Towards Seamless Interfaces between People, Bits, and Atoms. In *CHI'97 Proceedings.* Atlanta, GA.

Jenkins, H. 19??. 'No Matter How Small:' The Democratic Imagination of Dr. Seuss. To be published in H. Jenkins, T. McPherson, and J. Shatluck (Eds.), *Hop on Pop: the Politics and Pleasures of Popular Culture*. Durham: Duke University Press.

Jenkins, H. 1997a. Speech, Putting Boy Culture Back in the Home: Gender, Genre, and Virtual Play Spaces. *From Barbie to Mortal Kombat: A Conference on Gender and Computer Games.* Cambridge, MA.

Jenkins, H. 1997b. Unpublished lecture notes from *Problems in Cultural Interpretation: Reading Children's Fiction.* MIT undergraduate and graduate course, Spring 1997.

Johnson, W. 1996. *Penelope: the Story Weaver.* MIT Thesis, C.S. and E.E. Dept.

Kamptner, N. L. Personal Possessions and Their Meanings: A Life-Span Perspective. To Have Possessions [Special Issue]. *Journal of Social Behavior and Personality*, 6 (6), 209–228.

Kellner, D. 19??. Popular Culture and the Construction of Postmodern Identity.

Kincaid, J. 1992. The Child. In *Child-Loving: The Erotic Child and Victorian Culture*. New York: Routledge.

Kramarae, C. 1988. Gotta Go Myrtle, Technology's at the Door. In C. Kramarae (Ed.), *Technology and Women's Voices*. NY: Routledge & Kegan Paul.

Lewis, M.G. 1993. *Without a Word: Teaching Beyond Women's Silence.* New York: Routledge.

Lifton, R.J. 1993. *The Protean Self: Human Resilience in an Age of Fragmentation.* New York: BasicBooks.

Lipman-Blumen, J. 1984. *Gender Roles and Power.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Martlew, M. 1986. The Development of Written Language. In K. Durkin (Ed.), *Language Development in the School Years.* Cambridge: Brookline Books.

McLuhan, M. 1995. Understanding Media. In McLuhan, E. and Zingrone, F. (Eds), *Essential McLuhan.* New York: Basic Books.

Merriam-Webster Dictionary. 1996. Merriam-Webster, Inc.

Milne, A.A. 1996. *The Complete Tales and Poems of Winnie-the-Pooh.* NY, NY: Dutton Books.

Neruda, P. 1990. Towards an Impure Poetry. In B. Belitt (Trans.), *Selected Poems of Pablo Neruda.* NY, NY: Grove Press.

Norman, D. 1988. *The Design of Everyday Things.* NY: Doubleday.

Papert, S. 1996. *The Connected Family.* Atlanta, Georgia: Longstreet Press.

Peterson, C., and A. McCabe. 1996. Parental Scaffolding of Context in Children's Narratives. In C.E. Johnson & J.H.V Gilbert (Eds.), *Children's Language, Volume 9.* Mahwah, N.J.: Lawrence Erlbaum Asso.

Peterson, C., and A. McCabe. 1991. Getting the Story: A Longitudinal Study of Parental Styles in Eliciting Narratives and Developing Narrative Skill. In A. McCabe and C. Peterson (Eds.), *Developing Narrative Structure.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Poor, R. iRX 2.0 ... *Where Atoms meet Bits.* MIT Media Lab, World Wide Web document, 1996. http://ttt.www.media.mit.edu/pia/research/iRX2/index.html.

Poster, M. 1990. *The Mode of Information: Poststructuralism and Social Context.* Chicago: University of Chicago Press.

Resnick, M. 1994. *Turtles, Termites, and Traffic Jams: Explorations in Massively Parallel Microworlds.* Cambridge, MA: MIT Press.

Rose, J. 1984. Introduction *in The Case of Peter Pan: The Impossibility of Children's Fiction.* London: Macmillan.

Rostow Kuznets, L. 1994. *When Toys Come Alive: Narratives of Animation, Metamorphosis, and Development.* New Haven: Yale Univ. Press.

Sakai, M. 1997. *Storytelling Encouragement System*, MIT Master's Thesis, Department of Electrical Engineering and Computer Science, Cambridge, MA.

Stewart, S. 1993. *On Longing: Narratives on the Miniature, the Gigantic, the Souvenir, the Collection.* Durham and London: Duke University Press.

Strommen, E. 1998. When the Interface is a Talking Dinosaur: Learning Across Media with ActiMates Barney. Los Angeles, CA: *CHI '98 Conference Companion.*

Sutton-Smith, B. 1986. *Toys as Culture.* NY: Gardner Press.

Turkle, S. 1988. Computational Reticence: why women fear the intimate machine. In C. Kramarae (Ed.), *Technology and Women's Voices.* NY: Routledge & Kegan Paul.

Turkle, S. 1984. *The Second Self: Computers and the Human Spirit.* New York: Simon and Schuster.

Ullmer, B, and H. Ishii. 1997. The Meta-DESK: Models and Prototypes for Tangible User Interfaces. To appear in *UIST'97 Conference Proceedings.*

Umaschi, M. & J. Cassell. 1999. Children as Designers of Interactive Storytellers: "Let Me Tell You a Story About Myself... (This volume).

Umaschi, M. 1997. Soft Toys with Computer Hearts: Building Personal Storytelling Environments. In CHI *'97 Conference Proc.*, Formal Presentation and Demonstration, Atlanta, GA.

Williams, M. 1984. *The Velveteen Rabbit, or How Toys Become Real.* Philadelphia: Courage Books.

Winnicott, D.W. 1971. *Playing and Reality*. NY: Routledge.

CHAPTER THREE

Children as Designers of Interactive Storytellers

"Let me tell you a story about myself..."[1]

Marina Umaschi Bers and Justine Cassell

1. Introduction

In everyday life, conversational personal story telling is a communicative experience in which story listening is as important as story telling in order to construct a sense of identity. When people communicate they often do it through telling stories about their experiences and by finding personal relevance in other people's stories. New technologies have the potential to encourage people to share their personal stories, thus exploring their identity, as well as to learn about the nature of story telling interactions.

SAGE (Story telling Agent Generation Environment) is an authoring environment that encourages story telling and story making as well as provides a framework for children to design and program their own interactive storytellers. The latter supports changes in the way children think about themselves, and about story telling as a communicative activity.

Thus, SAGE supports two modes of interaction. In the first — story telling interaction — children are invited to share what is going on in their lives with a wise old sage, who "listens" and then offers a relevant traditional tale in response. In the second — storyteller authoring interaction— children are invited to add to the library of wise old sages by designing their own storyteller for themselves and other children to interact with. In order to support children as designers as well as users of the storytellers, we imple-

mented a visual authoring language that allows children to program (1) the scripts that are told by the storyteller, (2) the conversational structure, or flow of the interaction between user and storyteller, (3) the body behaviors of the interactive stuffed animal —the storyteller's assistant, (4) the database of stories that are offered as the storyteller's response.

2. Context of the Work

Although little research has been done on how technology can be used to encourage children to learn about identity and communication through story telling, a fair amount of research exists in relevant fields that have an impact upon these issues. In what follows we present a theoretical framework, drawing from research in diverse areas such as narrative theory, psychology, artificial intelligence, education, and human-computer interaction. We are aware that a complete survey of these research areas is impossible, therefore we chose to focus on the particular aspects that guided our system design.

2.1 *Telling and listening to stories*

Narrative is the primary form through which we understand and give meaning to our experience (Polkinghorne 1988). Our research focuses on a particular kind of narrative — conversational stories of personal experience — defined by Polanyi (1989) as highly structured linguistic productions found in the context of everyday talk, and by Miller et al. (1990) as face-to-face interactions in which self expression is accomplished by the recounting of personal experience.

Narrative, including conversational stories of personal experience, serves at least three vital functions that were carefully analyzed in order to design the SAGE system.

1. *A cognitive function.* Personal stories are fundamental constituents of human memory, and new experiences are interpreted in terms of old stories and generalized story scripts (Bartlett 1932; Schank & Abelson 1995). Bruner (1986) describes narrative as a mode of cognitive functioning that provides a distinctive way of ordering and understanding experience.
2. *A social function.* The tales that one knows and can tell define the social group or culture to which one belongs (Turner 1980). Myths, legends,

and traditional tales provide a sense of continuity between generations as well as models for human behavior (Campbell 1988). Conversational personal stories play an important role in the social construction of the self from early childhood (Miller et al. 1990). Adults tell and retell life stories according to certain conventions and in many different contexts as a way of establishing coherence in their lives (Linde 1993).

3. *An emotional function*. Story telling has been used in very different forms of psychotherapy (Wigren 1994). In fact, one view of psychotherapy is that it leads us to be able to tell coherent life stories (Polkinghorne 1988). The emotional value of tales has been explored by Erickson in hypnotherapy (Rosen 1982) and in fairy tales by Bettelheim (1976). As Anna Freud (1965), Erik Erikson (1950) and others have shown, through the verbal-play experience of story telling, children can find not only recreation but also self-cure.

From cognitive, social and emotional standpoints it is important for children and adults to have a place to tell their story. But it is also important to provide a space for them to explore what is narrative and how conversational story telling works. SAGE was conceived to support both story telling and story making. While story telling helps children to "play out" what is happening in their lives, story making supports learning about narrative by interacting with and designing meaningful storytellers. Taken together, these two activities allow users to construct and express their narrative voice (Cassell 1998) as well as their identity.

2.2 *Story telling systems: beyond story writing*

To date most of the research on computers and story telling has focused on interactive games, mystery simulations, and interactive fiction (although see Don 1990 and Laurel 1993). Interactive games and mystery simulations present a plot by way of short descriptions of characters and places, and they allow users to navigate through the spaces and find out more about the characters. Unlike narratives about the self, these are "forms of narrative that privilege space over characterization or plot development. [They are] part of an alternative tradition of 'spatial stories,' a different way of organizing narratives that must be examined and evaluated according to their own cultural logic" (Fuller & Jenkins 1995).

Interactive fiction presents a short story or novel in a hypertext format,

the branching nodes allowing users to read a different story each time by choosing different links. Unlike personal narratives, this work concentrates on how to adapt the genre of fiction to the constraints and possibilities of the computer. Other work has used computation to assist production and viewing of narrative structure models for cinema. For example, "Agent Stories" (Brooks 1996) is an environment for non-linear, cinematic story design and presentation that can be used by cinematic story writers.

In the entertainment domain, the design of story telling systems for children has focused on interactive games and fiction of the type described above. In the educational domain, work has been done on supporting children's story writing process. Commercial software such as Kid Works Deluxe (Davidson 1995) and the Amazing Writing Machine (Brøderbund 1994) are augmented word processors that encourage children to tell stories. Educational researchers have also devoted extensive effort in developing story writing software to aid college essay writers in brainstorming and planning (Daiute 1985; Burns 1984). At the MIT Media Laboratory, Montford (1998) designed EddieEdit, a conversational computer character to aid children in the process of writing stories, with special focus in planning and revision.

By contrast, the research presented in this paper is not aimed at improving story writing, rather SAGE seeks to support children's authoring of their own wise storytellers and story listeners through the design of an interactive character. The authoring process involves the design and implementation of conversational interactions between the user and the wise storyteller. This serves as a vehicle for the transformation of ways of thinking about narrative as a communication process deeply involved in identity construction.

SAGE employs the notion of a simple conversational system with a well-defined character, similar to Eliza (Weizenbaum 1976), Parry (Colby 1975), and Julia (Mauldin 1994). These are described by Murray (1991) as parodic interactive character whose computational rigidities model recognizably human types. SAGE extends the domain to the world of narrative, and allows storyteller construction as well as interaction. It has been suggested that the construction of emotionally believable characters with a limited field of interaction can maintain the suspension of disbelief of the user and thus allow successful human-computer interaction in the absence of deep Artificial Intelligence (AI) (Bates et al. 1995). This alternative to the deep AI approach focuses on believability over ability.

By "ability" we refer to the vision that computers should try to re-create the cognitive processes that people use to understand stories — as a step

towards machine understanding with broad applicability. For example, there is extensive research on producing models of the world that contain particular knowledge organized around standard situations, scripts or cases (Schank & Riesbeck 1981). This story understanding research depends on a theory of narrative as primarily constituted by goals and actions in the context of plans. A different tradition within AI, drawing from anthropology and ethnomethodology instead of cognitive science, has proposed a theory of situated action as complementary to the traditional planning approach (Suchman 1987). The emphasis is not on mental processes but on social relations produced through the actions people take in the world. The integration of social and cognitive approaches has been proposed by understanding how meaning involves a social as well as a mental dimension (Winograd & Flores 1986). These theories are compatible with SAGE's notion of stories as social as well as cognitive artifacts.

In SAGE, a strong sense of context and a shared assumption about the socio-cultural role of the storyteller allow the user (1) to believe in the abilities of the system, despite minimal natural language understanding, and (2) to find coherence between his or her personal stories and the response stories retrieved by the system. This is achieved by enabling children to design storytellers with strong stereotypical characteristics and well-defined domains that set up certain behavioral expectations (Foner 1998).

2.3 *Learning by designing*

SAGE was specifically designed as a computational construction kit (Resnick et al. 1996) to help children learn in an active and hands-on way. Its educational philosophy is based on constructionism (Papert 1980) which asserts that learners are likely to have access to different kinds of new ideas when they are building artifacts that they can reflect upon and share with others in their learning community. SAGE is a construction kit that supports two types of connections required by constructionism to bootstrap learning:

- personal connections — children can program their interactive embodied storytellers according to their interests.
- epistemological connections —new ways of thinking about story telling, identity and technology are encouraged.

In the light of this approach, SAGE focuses on the building of stories and sage storytellers as "evocative objects" (Turkle 1984 & Turkle 1995) that invite

reflection about identity and communication. SAGE provides an authoring environment and programming language where children can create their own interactive storytellers. When children design storytellers and describe the underlying turn-taking rules and body movements of the storyteller's toy assistant, they behave as communication and narrative theorists as well as programmers. In the process of creating their own simulated characters they become explicitly aware of the structure of the conversation; for example, in order to allow turn-taking between user and system, they must understand the nature of conversational units. Programming in this case is a vehicle for the transformation of ways of thinking about the self and communication as well as for constructing knowledge about technology.

2.4 *Computationally-augmented soft toys for story telling*

SAGE embeds the sage storyteller's assistant in a programmable interactive stuffed animal — a soft interface (see Figure 1) — that carries on the conversation with the user. The stuffed animal is capable of some of the types of nonverbal behaviors that humans use to indicate engagement, and that are commonly found in conversational narratives between people (Cassell 1995).



Figure 1. Interactive programmable stuffed rabbit and Handy Board (microcontroller). (Copyright ©1998 by ACM)

In interaction mode, children can watch the stuffed animal move its ears, shake its body and blink its eyes to indicate attentiveness as they converse with it. In design mode, children can program the toy's communicative behaviors by specifying an action that the toy would perform at each turn in the conversation. They can also create the different personalities and conversational styles that the toy might demonstrate.

Stuffed animals are objects whose emotional affordances can be exploited in building computational interfaces for children (Druin 1987; Umaschi 1997a; Glos 1997). For example, Microsoft has developed an interactive toy called Barney. The toy communicates with the PC wirelessly, reacts to CD-ROM games and watches television with the child (Alexander & Strommen 1998). The psychologist Winnicott (1971) introduced the term "transitional object" to refer to the first "not-me" infant's possession — often a stuffed animal. The favorite toy, as described by Winnicott, "must seem to the infant to give warmth, or to move, or to have texture, or to do something that seems to show it has vitality or reality of its own."

Research on human-computer interaction has taken on challenge to unite the digital with the physical world so as to integrate the power of computing seamlessly into toys. This area is moving from graphical user interfaces (GUI) to tangible user interfaces (TUI) — from a mouse and keyboard to ubiquitous computing and tangible media (Ishii & Ullmer 1997). An example in the domain of technology for children is Dr. LegoHead (Borovoy 1996). In this system physical Lego blocks have been augmented with digital information, allowing children to build creatures out of high level Lego brick parts such as eyes and mouths, thus exploring ideas about the nature of metaphor and function. Following this line and in order to support emotional engagement, SAGE offers the possibility to embed the wise storyteller's assistant into a programmable stuffed animal with basic motor functions.

SAGE takes the challenge of designing tangible user interfaces and leverages from the fact that children establish intimate communicative relationships with their stuffed toys (Dyl & Wapner 1996). On one hand, the natural tendency to communicate at a deep level with soft objects can be leveraged by adding computational abilities so toys can give feedback. On the other hand, children can program the toy's communicative behaviors, thus exploring powerful ideas about the nature of communication. The result is computationally-augmented soft toys for story telling.

3. SAGE: Design and Implementation

SAGE was designed to support the creation of wise storytellers who have a comforting story at hand in response to the user's experiences. Children can engage with SAGE in two modes: (1) by interacting with a sage storyteller chosen from a library of already existing characters, and (2) by creating their own sages and types of story telling interactions. The LISP-based SAGE architecture has three modules described at length in the following subsections:

- SAGE computation module: in charge of parsing the user's story, expanding keywords through WordNet and performing a match between the user's personal story and a comforting story in the database.
- SAGE authoring language: used by children to design their own interactive characters, the conversational flow between user and storyteller and the data-base of stories offered by the system in response to the user's story.
- SAGE interface: the interactive toy and its motor behaviors, computer screen and sound (output), and the keyboard (input).

3.1 *The computation module*

The SAGE computation module is in charge of processing the user's story and retrieving a comforting story for the sage storyteller to give in reply. First, the user's story is parsed. Nouns and verbs are extracted and submitted to WordNet, a hierarchical semantic lexical reference system in which nouns, verbs and adjectives are organized into sets that represent concepts (Miller et al. 1993). WordNet returns a list of synonyms, hyponyms, and hypernyms of the original keywords. Different keywords that can be grouped under one single hypernym result in weighting that concept more highly than other ones.

The final list of weighted nouns and verbs is a way of representing the user's story at a more abstract level. The weighted nouns and verbs are matched against nouns and verbs indexing stories in the story database. The story that is most like the user's story — that deals with the same themes — is retrieved and told to the user. This computational module does not have any knowledge about story grammar, it only deals with augmented keywords. It presents a simple implementation solution to the complex problem of parsing and indexing stories about personal lives (Domeshek 1992).

3.2 *The authoring language module*

The SAGE authoring language has a graphical user interface that allows children both to design storytellers to interact with and also to model different types of story telling situations. The environment supports children's creation of interactive storytellers by providing three design windows that can be chosen from a menu: the character's creation window, the conversational structure window and the database of stories window.

In the character's creation window (see Figure 2) children describe different facts about the lives of the sage storyteller and an optional storyteller's assistant — the interactive toy. These facts will determine how the storyteller (and assistant) interact with the user. This window is also used to load pictures of the storyteller who lives on the screen (scanned photos or sketches done in a drawing program).

In the conversational structure window children design the conversational flow between user and storyteller (or the storyteller's intermediary, the toy assistant) by selecting and arranging objects of three different types from a palette — turn-taking states, communicative actions, and parts of conversation. These objects are placed by the user in the conversational structure window in order to model the conversational story telling interaction (see Figure 3). Once these objects have been selected, the scripts that instantiate

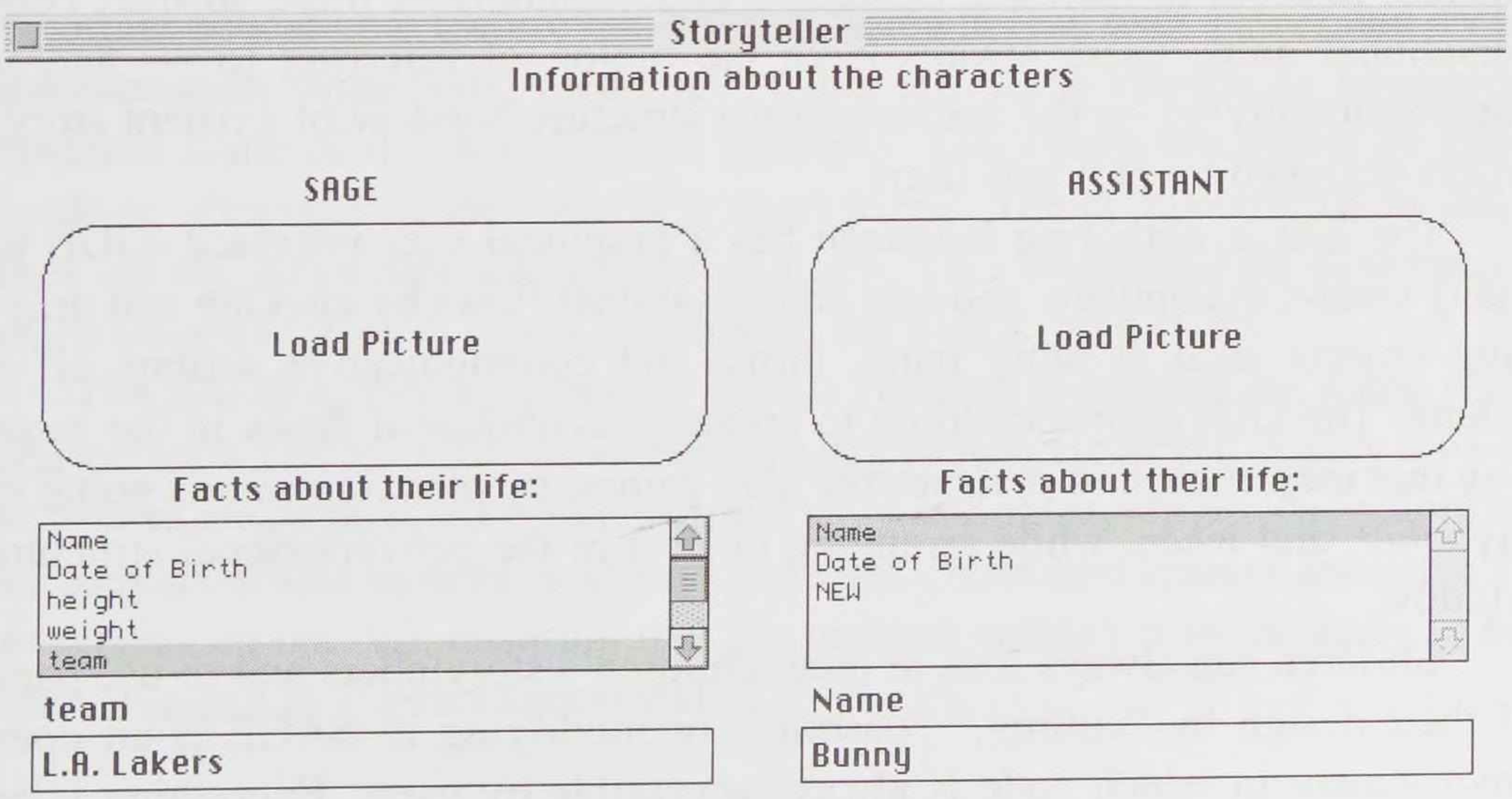


Figure 2. Window to design character's personality

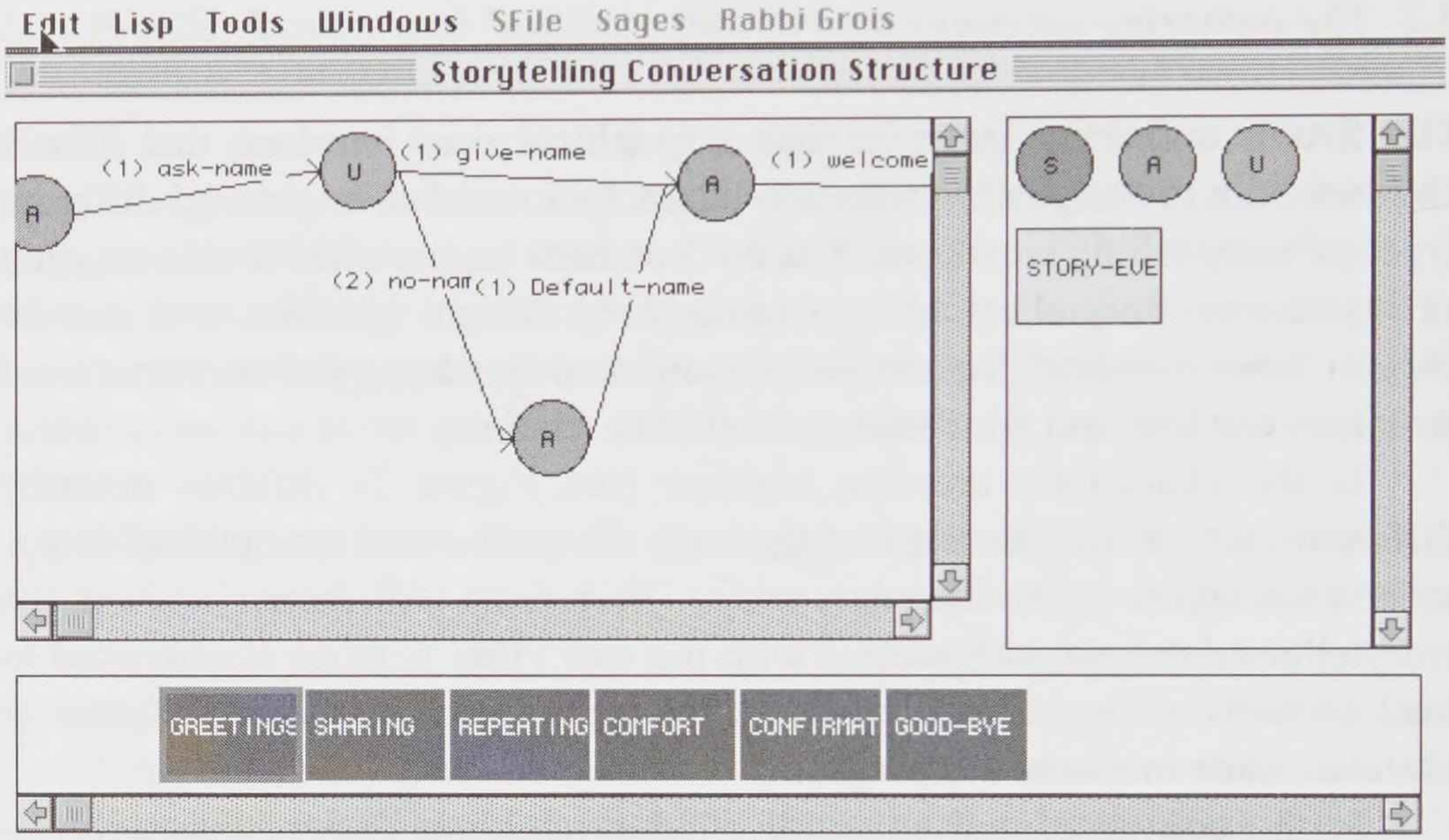


Figure 3. The conversational flow screen is divided in three. On the right, the palette to select parts of conversation and turns: user (U), interactive toy assistant (A) and storyteller (S). On the left, the space to combine turns (circles) with communicative actions (arrows). In the bottom half is the plot of parts of conversation.

them can be written in the window that pops-up when the object is double-clicked. In order to facilitate children's understanding of these abstract conversational units, users always have the option of referring to — and/or copying/modifying — the conversational structure window of existent storytellers designed by previous users.

The SAGE authoring language has a graphical user interface (GUI) to easily create, manipulate, and edit conversational flows by clicking and dragging objects such as story parts, turns, and communicative actions off a palette. The GUI allows children to create conversational flows in the same way that they engage in pretend role play games, by planning *who* is going to say *what*, and *when*, while arranging objects in the conversational structure window.

Children can always look at other children's storytellers and re-use parts of their design by "cutting," "pasting," or modifying it. SAGE is an open environment in which code is always accessible by users. Everything is an example for others to use.

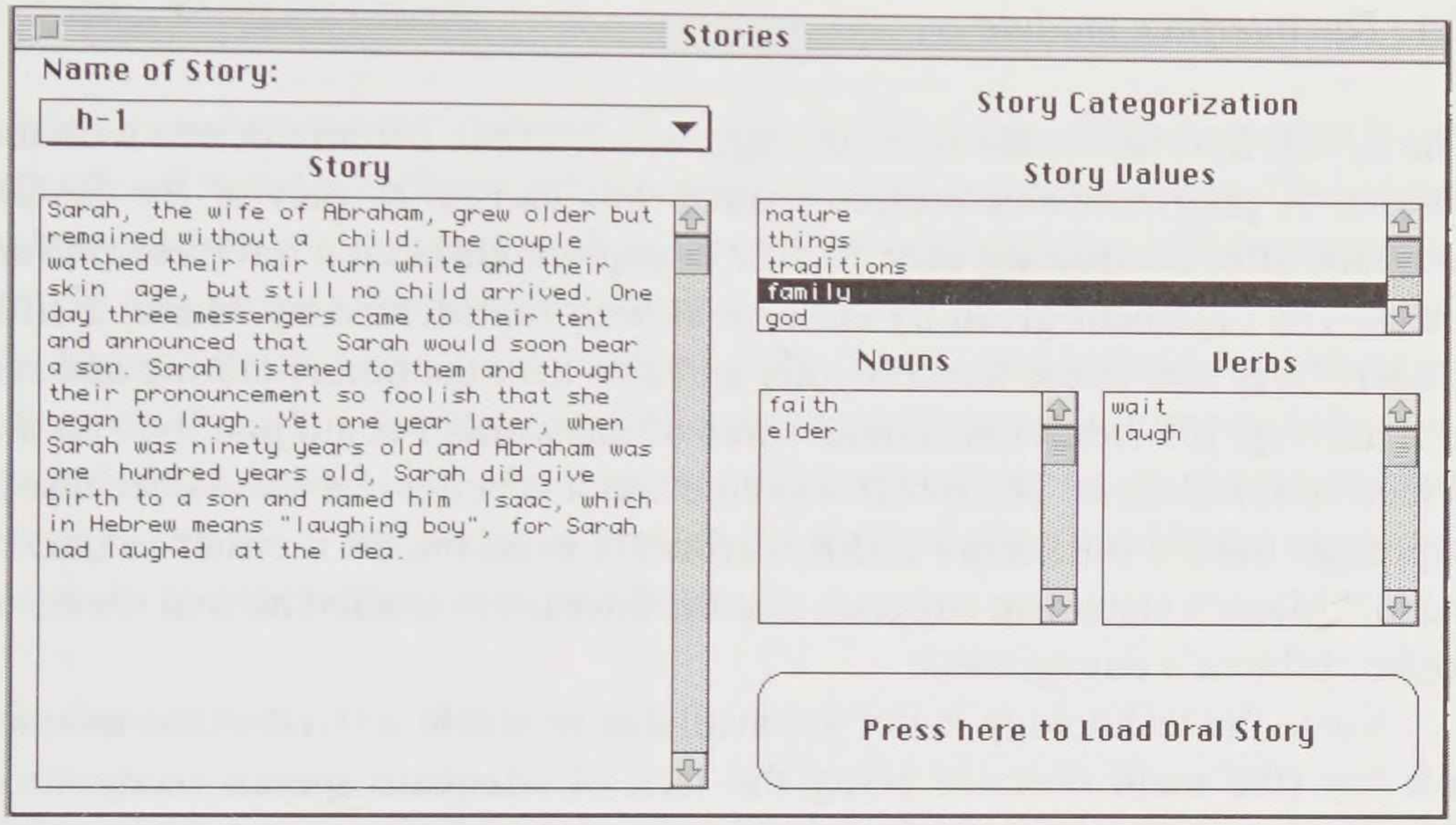


Figure 4. The database of stories and its descriptors

In the database of stories users can write or scan in response stories — the comforting stories that will be offered by the sage. They can also record them in their own voice or ask someone else to do it. Children categorize or annotate the stories with story values, nouns and verbs that they consider good descriptors of the main story points (see Figure 4). The hand annotation process is an important educational feature that helps children to read over a story and try to understand its underlying meaning. Nouns are verbs chosen as keywords represent some of the story's basic concepts. The values can either be keywords or sentences in the form of a short moral. The computational module takes all the keywords and augments them. This augmentation determines which response story is a best match with the user's input story.

The SAGE authoring language allows children to quickly get going with their projects and to create an interactive storyteller. However, it only allows the user to create directed conversational story telling situations. The nature of the conversational system is always machine-controlled (mixed-initiative is not readily an option), ensuring that the persona created is more likely to be believable and to interact successfully.

3.3 *The interface module*

The SAGE interface is the layer through which a user communicates with the system. A programmable stuffed animal was chosen as part of the SAGE interface after conducting user studies to explore children's interface preferences. The argument given by children is well represented by Marie, a fifth grader: "It is neat to use the computer and also hear the bunny talking and see him moving. If I didn't understand some of the words I could just look on the screen and read them. The rabbit was cute and it is better if it moves because it expresses itself a little more and sometimes it even makes it easier to understand." Marie's statement suggests that the interactive stuffed animal seems to foster children's engagement.

Since the technology is not yet available to create a stand-alone interactive toy (the main obstacle being the lack of adequate speech recognition systems for children's speech), we integrated the desktop computer with a physical toy. Output is multimodal, and input is typed text only. In order to explain the absence of symmetry in the interaction, the stuffed animal is presented as being the assistant of the sage storyteller that lives on the computer screen (see Figure 5). The toy is in charge of carrying the conversation with the user and, at a certain point, calls on the wise storyteller who lives on the screen to tell a traditional tale of wisdom.

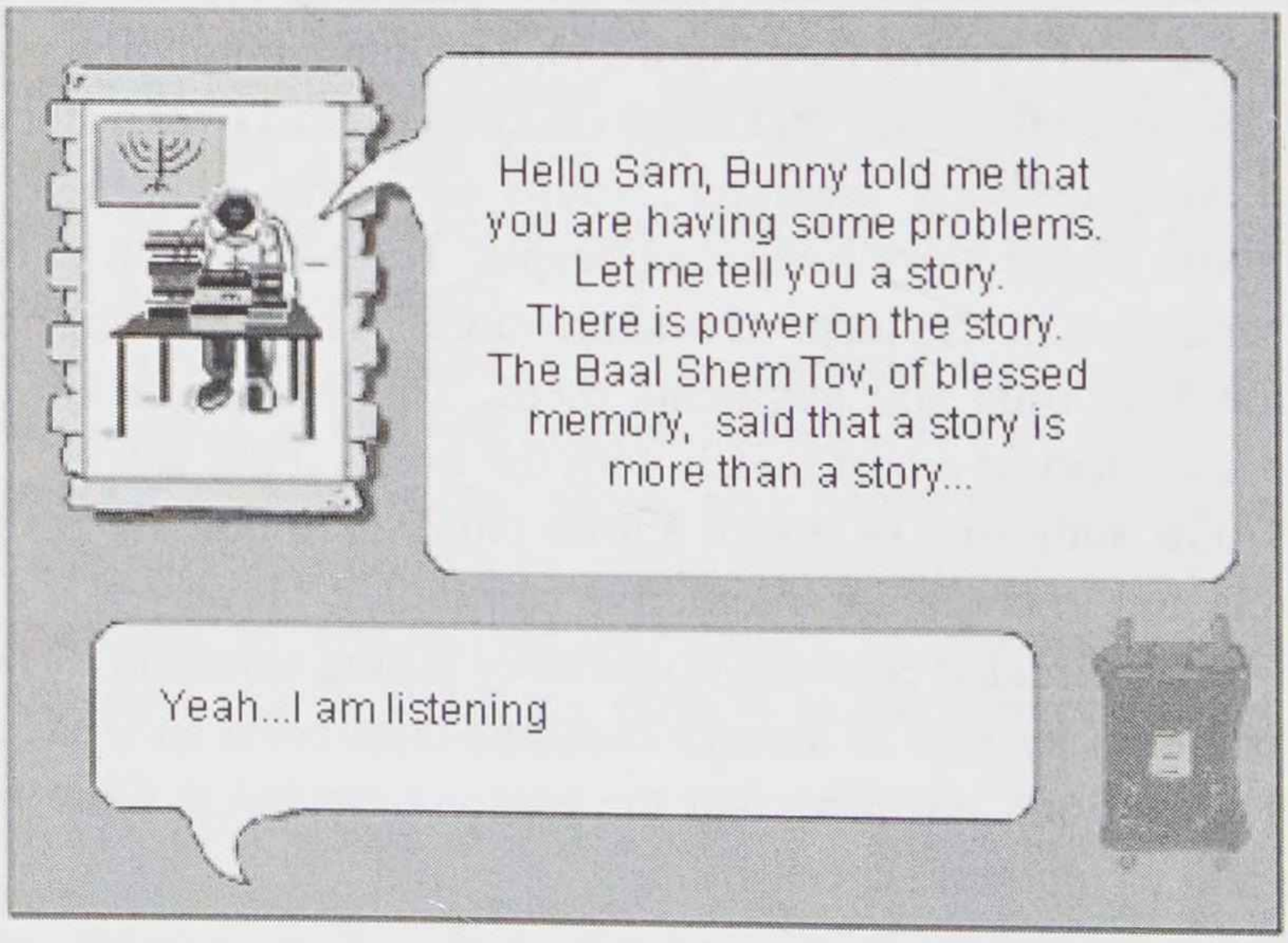


Figure 5. The sage storyteller, in this case a Hasidic Rabbi, lives in the screen and is in charge of telling the response stories

4. Empirical Research

In the introduction we claimed that the SAGE story telling system is an example of how technology can be used to encourage children to reflect about their inner life, and to support their learning about narrative and communication. In this section we describe empirical research that supports these claims. We show that children's interactions with SAGE did indeed lead them to reflect on thoughts, feelings, and interpersonal communication, and also expanded their knowledge about personal story telling as a communicative activity. The following sections present the methodology used in the empirical research, and then address these claims.

4.1 *Description of workshops*

We conducted several pilot studies to examine children's interactions with the sage storytellers, and their ability to build their own meaningful characters. Our earliest research showed that children were in fact quite ready to engage deeply with sages that we had designed (Umaschi 1996). They opened up willingly to tell personal problems to the characters that we had designed, a Hasidic Rabbi and a Buddhist scholar. Here is an early sample of a conversation that a thirteen-year old had with the Hasidic Rabbi:

SAGE: You seem a little shy, Leo. I also was shy to tell stories, but since my father and my grandfather were Hasidic Rabbis I knew I had to become one. That is the way it is. Nu,Leo how old are you?

USER: 13

SAGE: Oy vey, I am 77 years older than you. You know Leo, a story is more than a story...a story has power. Nu, what is your story?

USER: I had my first day at school today and everybody was making fun of me because I did not know the Guemarah(sig)

SAGE: Oy vey... go on, Leo.

USER: so they started to beat me up because I was so stupid

SAGE: I understand — what more can you tell me Leo?

USER: and everyday when I come home from school I cry for a long time

SAGE: Nu, do you think you are the first one in this situation? Leo let me tell you a story: When Rabbi Wolf Zbarazer was on a journey, he would not permit the driver to beat his horses, saying to him: "You have no need even to scold them, if you understand how to address them properly." Can you find a connection with your own story?

The nature of children's engagement was no different whether they under-

stood the workings of the implementation or believed that the computer system learned something deep about them. This result led us to believe that children would also be able to interact with sages of their own design.

In the study described here, we conducted research with fourth and fifth graders in order to test SAGE's impact on story telling awareness and self-reflection skills, and the interaction between these skills and technological fluency (Papert & Resnick 1995). Children of this age are almost adult-like in their narrative productions (Karmiloff-Smith 1985; Hickmann 1987) but are still not capable of explicit meta-linguistic comments about the structure of narrative (Gombert 1992). Likewise, American fifth graders are in general familiar with computer applications but for the most part are not fluent programmers. The general claim here is that SAGE integrates programming and story telling skills by supporting the design and modeling of abstract structures needed and used in both activities.

In order to understand children's design and creation of their own meaningful storytellers, we conducted two-day weekend workshops. Children were solicited for the workshops through local schools and personal connections. Children who participated in the workshops came from a range of backgrounds, and were not particularly computer-competent. Individual extended interviews were carried out before and after the workshops to explore the children's ability to be self-reflective, their notions of story telling, as well as their computer skills. Eight children (four boys and four girls) participated in the study reported here. Each workshop comprised one pre-study personal interview, two eight hours days of design and programming, one post-study personal interview, and one demonstration for parents given by the children. Each child worked on a storyteller of his or her choice, and kept his or her own design notebook with ideas and problems. During the design process children interacted with each other, asked questions about their programs and played and helped debug each other's storytellers.

The methodology used for the study was based on an ethnographic approach, with natural observation of the children's interaction with the system as well as with each other. Analysis of system logs generated by interaction with SAGE, video-taping, extended personal interviews and an experimental task also took place (for a complete description of the methodology and the questionnaires used in the user studies, please see the appendix in Umaschi 1997b). In order to avoid "technocentric questions" (Papert 1987), the methodology was centered on what children do with SAGE and not what

SAGE does to children. In the following sections we look at examples of wise storytellers created by children. Each example is a case study that illustrates a phenomenon also found more generally during the workshops.

4.2 *The presentation of the self*

When interacting with one of the sages that we had designed, children revealed aspects of their inner lives and the problems they face everyday. In building a sage, children designed that person to whom they wished they could turn with their problems. But it is essential to note that the sages they designed were also aspects of the inner life of the child designer. That is, since the children were creating their own characters and databases of stories, they were also playing with different notions of self, and creating or imitating the narrative voices they wanted or needed to hear. Children created storytellers as projections of fears, feelings, interests, and role models. These projections allowed the presentation of the self to themselves as well as to others.

Out of a total of eight storytellers built during the workshops, three were based on humans or fictional media characters: Doug, a cartoon character; Tera Randof, based on a mixture of a real sports figure and family members of the child designer; and Shaquille O'Neil, a well-known basketball player. The other five characters were fictional, although two of them, Spot and the Big Orange Fox, clearly tell stories based on the designer's personal experiences.

Abi is a bilingual fourth-grader whose parents come from Russia. When he grows up he wants to be a physicist because he likes computers, math and science. Unlike all the other children who participated in the workshop, he said that he had never heard the word identity before, and after we explained its meaning he concluded: "it is what is inside me, like being Jewish and American and having my relatives in Russia."

During the workshop Abi created The Big Orange Fox that would explicitly ask the user if she or he has specific problems. If the user responded positively, The Big Orange Fox would tell an appropriate story indexed in the database under different types of problems. For example, below is a story written by Abi and told by his character, The Big Orange Fox. Abi indexed this story as religion problem.

> There was once a boy who did not like his religion so he followed his family religion but when he was alone he secretly followed his religion but when his parents spotted him they cut off his finger and put him into a tiny dark box with spikes on all the sides but the bottom and left him there for 24 hours.

Abi wrote twelve stories of this sort in his database. He is the child who created the most number of stories. This is extremely noteworthy considering that in the pre-interview Abi stated very clearly that he didn't like writing and he only liked to do it with a partner, not alone.

During the demonstration day, when Abi's parents came to play with The Big Orange Fox, they were surprised by the questions the character asked but even more by the stories it offered. While playing with other children's characters they noticed that Abi's was addressing in a direct way personal problems as well as telling stories, that although fictional, were obviously expressions of fears and feelings.

During the post-interview, Abi said:

> First of all I learned how to put myself in another place and pretend I am someone else; and I learned how hard it is [...]. But really I talked to myself and I learned more about myself [...] My stories are sad, but if you want to hear a funny story you can play with Waloompa the alien. When you have a problem it is sad and that is why my stories are sad. But my family didn't like my stories too much because of that; but they learned on what place I am now, and I am in the sad stories.

Abi used SAGE to create a fictional representation of himself, for himself and for his parents. His parents discovered, through The Big Orange Fox, a part of Abi's inner world.

In the same way that Abi used SAGE to play with his fears and feelings, Tracy, a ten-year-old girl with no previous programming experience, used it to represent herself in the future. During the workshops Tracy created Tera Randof, a young female basketball player "who will become a sports figure" like the ones Tracy admires. In the post-interview, Tracy revealed that Tera's personality is a mixture of Tracy's mother and grandmother, and that the stories that Tera tells are Tracy's mother's childhood stories. Tracy created the character and the conversational structure, and wrote all the stories in the database. By interacting with her own storyteller over and over again, Tracy improved the program until it was able to respond to many different kinds of user's stories without crashing. In the process, Tracy also reflected on how people become good listeners. And, despite the fact that she was interacting with a storyteller of her own design, Tracy shared real personal problems and deeply engaged in reflection to make sense of the given advice. The following interaction between Tracy and her sage storyteller took place during the design process:

SAGE : I have a lot of problems like wondering if my team is going to win the big game or if I will mess it up for everyone. Please tell me a problem you have.

USER : I am sad because my parents told me that we can't take my dog on vacations with us, so I will miss him.

SAGE : Your problem reminded me of this story. Kerri is my best friend. Kerri knows everything about me and I know everything about her. Every Friday we sleep over each other houses. Switching houses each week. This week it was at my house. I got the usual snacks: popcorn, pizza, chips, soda, and candy. A few hours later Kerri came by with her sleeping bag. You could tell she was in a bad mood. I asked her to lighten up, and she snapped, "Fine" and left. I was stunned. What was her problem? I waited the few minutes that it took for her to walk home then called her. Her mom answered. I asked to talk to Kerri. Her mom said Kerri doesn't feel like talking to you right now. Kerri didn't talk to me in school for a month. Then I got the horrible news: my mom got a great job in L.A. so we were moving there! Kerri didn't even come over to say good-bye. I have never seen or talked to her since. Maybe she wasn't really my best friend. Did you like that story? Did you understand the moral of it? Please tell me what you think that story meant.

USER : You didn't see your friend ever again, but I will see my dog when we come back from vacation. I guess that is not that bad.

Tracy is only one of the many cases in which we observed, by analyzing the logs of the interactions, that a deep knowledge of how the system works does not affect emotional engagement and sharing of a personal problem. Both Abi's and Tracy's case studies show that SAGE was effectively used by children to explore identity, according to their personal needs at the time.

4.3 *Powerful ideas about communication*

Children's knowledge of how SAGE works allowed them to reflect about the nature of sharing problems in a deeper way and compare it with their intuitions about how people and machines communicate with each other. Tracy said:

> A person already knows how to react to a question, but the computer does not. You have to feed in information about what to say and that is like creating a mind and a personality [. . .] When people communicate they take an experience they had or think of something that the other person might want to hear, a same sort of problem, something familiar or similar, and they just tell it.

Mischa, created Mother Nature and designed a complex conversational structure that has two parallel branches, one for people who like to be in nature and one for people who do not. In the post-interview, reflecting on her authoring experience, Mischa said,

> The most fun part was working on the computer and programming kind of a game for people to use [...] it is a program for you to make another program. I liked creating what it might say and the questions and the structure and I learned that people can type anything and that you always have to be prepared because people might type not just "yes" or "no" but, for example "I guess so", so you have to use more branches.

Both Tracy and Mischa concluded that one has to be ready with an appropriate output for any input. This is a powerful idea for a young child to start thinking about what is communication. SAGE facilitates the exploration of notions of communication as children observe conversational breakdowns in their own and others' interactions with the created characters. The most frequently observed communication breakdown was due to children's difficulty in decentering and designing interactions in which other people, different from themselves, could participate. By 'decentering,' we mean the ability to imagine a conversational interaction from the point of view of another interlocutor; that is, more generally, the moving out of the phase of egocentrism where one cannot differentiate somebody else's point of view from one's own.

The iterative design experience, which involved programming and immediate testing with a real audience — themselves, their parents, and other children — gave children the opportunity to decenter and debug their conversational structures after observing other people's interactions. Bernie's experience building his idol Shaquille O'Neil, the basketball player, is a good example of the difficulty children had in decentering.

When Bernie interacted with Shaq, the program responded well. However, when someone else tried it, the first thing that happened was that Shaq called them "Bernie." Bernie realized that he had hard-coded his name in all the system responses. The problem was not a computational bug, but a communication bug. Bernie had not decentered, i.e., he had not created an experience for users different from himself.

A week later, when his parents came for a demonstration, Bernie had already fixed "the name problem, " as he called it. However, when Shaq asked Bernie's dad, "I guess you are around 11, am I right?" and the response was, "No, I am 45," Shaq crashed. This time Bernie saw the problem immediately. He hadn't provided a branching node to allow the possibility of a "no." He went back to the authoring mode and added the branching to the conversational structure.

Bernie's experience was common to most of the children, who observed

similar problems when their parents interacted with their characters. In general, the conversational structures designed were very linear, without branches to take into account different kinds of user input. Only after observing others' interactions with their sages did the children realize that they had to contemplate the possibility of responses other than their own. This discovery was facilitated by the iterative design process, meaning the loop of working on the computer and immediately debugging or trying results in the real world in an iterative process.

5. Conclusions

In this chapter we discussed the importance of technological tools specifically designed to encourage children's exploration of identity and communication. We proposed personal story telling as a way to access the inner world as well as to explore notions of communication. Constructionism informed the educational philosophy of our work.

We presented the design and implementation of such a technological tool: SAGE, an authoring environment for creating embodied interactive storytellers with whom one can converse. We described children's interactions with the SAGE tool, concentrating on case studies.

Contributions of this research are the themes that emerged from the analysis of the children's learning experiences:

- The continuum between the interacting and the authoring mode: From the self-reflection perspective, both experiences were successful. However, the authoring mode seemed to support deeper explorations of identity and communication as well as a bigger sense of empowerment and control over the learning experience.
- The creation of storytellers as projections of fears, feelings, interests and role models: These projections allow the presentation of the self to ourselves as well as to others.
- The exploration of notions of communication while observing breakdowns: The iterative design experience supported this type of exploration by providing children with the opportunity to decenter and debug their system's conversational structures after observing other people's interactions with them.

This work provides a conceptual and technological framework for designing and implementing tools for reflecting about the inner world as well as exploring communication issues. Both activities can also be done without technology, through mentoring with a good counselor or teacher. However, the use of computers enables a different exploratory depth as well as supports children's experiential learning while they are modeling and designing abstract conversational story telling structures. The endeavor is just beginning, and further research must be done in order to explore in depth each of the themes that has emerged from this work.

For the moment we have just started to examine the interaction between technological and story telling fluency across single weekend workshops. In the future we hope to look at the effects of interaction with SAGE over a longer period of time. We have also begun to look at other kinds of functions that interaction with a sage storyteller can sustain. Some children are particularly in need of telling the stories of their lives — for example those children whose stories are complex and emotionally wrenching, such as chronically ill children (Bers et al. 1998). SAGE has the characteristics of a gender neutral technology which could support girls to become familiar with the world of technology and boys to become more comfortable with story telling. However, gender-based empirical research needs to be done in order to test this hypothesis.

We live surrounded by stories and by technology, sometimes in combination. We do not always need to understand the structure of a story to enjoy it; however, we do need to be able to recognize something about the structure in order to become critical readers of our reality. Our hope is that by supporting children's creation of their own story telling structures, children will become more aware of the metacommunication cues that signal the goals or agenda with which certain stories are told or written.

We also live among technology and increasingly technological fluency is demanded of us in all aspects of our daily lives. Our hope is that by supporting children's design and programming of their own meaningful projects, they will grow up with a sense of control and empowerment. We believe that technological tools for reflecting on the inner world do not only improve children's sense of themselves, but that children's sense of themselves will improve the technology around us.

Acknowledgments

Thanks to the members of the Gesture and Narrative Language group and the Epistemology and Learning group at the MIT Media Laboratory. We are thankful to Prof. Mitchel Resnick, Prof. Sherry Turkle, Prof. Seymour Papert and Edith Ackermann for insightful discussions. We also thank Dr. Fred Martin and Rick Borovoy for providing support with the handy-board, and Warren Sack for introducing us to WordNet. Special thanks to the Interval Research Corporation for supporting part of this research and to Lee Felsenstein and his team for building the interactive rabbit. We are also grateful to the crew of talented undergraduates, Ien Cheng, Adrian Banard, Peter Davis, Anthony Young-Garner, Andy Buttner and Yaron Koren who worked on different aspects of the implementation of SAGE. Big thanks to Josh Bers for reading and correcting many versions of this manuscript.

Notes

1. An extended version of this article, "Interactive Storytelling Systems for Children: Using Technology to Explore Language and Identity", was published by AACE in the *Journal of Interactive Learning Research*, volume 9 (2), copyright 1998.

References

Alexander, K & E. Strommen. 1998. Evolution of the Talking Dinosaur: The (Not So) Natural History of a New Interface for Children. In *CHI'98 Proceedings*, ACM, pp. 7–8.

Bartlett, F.C. 1932. *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Bates, J., P. Maes, & B. Hayes-Roth. 1995. AAAI Working Notes Spring Symposium: Interactive Story Systems: Plot and Character. Presented at AAAI Conference, Stanford University.

Bers, M., E. Ackermann, J. Cassell, B. Donegan, J. Gonzalez-Heydrich, D. DeMaso, C. Strohecker, S. Lualdi, D. Bromley, & J. Karlin. 1998. Interactive Storytelling Environments: Coping with Cardiac Illness at Boston's Children's Hospital. In *CHI'98 Proceedings*, ACM, pp. 603–609.

Bettelheim, B. 1976. *The uses of enchantment : the meaning and importance of fairy tales.* NY: Knopf.

Borovoy, R. 1996. *Genuine Object Oriented Programming*. SM Thesis. MIT, Cambridge, MA.

Brooks, K. M. 1996. Do Story Agents Use Rocking Chairs? The Theory and Implementation of One Model for Computational Narrative. In W. Hall & T.D.C. Little (Ed.), *ACM Multimedia*, pp. 317–328, Boston, MA.

Bruner, J. 1986. Two modes of Thought. In *Actual Minds, possible worlds.* Cambridge: Harvard Univ. Press.

Brøderbund. 1994. The Amazing Writing Machine. CD-ROM for Windows/Macintosh.
Burn, H. 1984. Recollections of First-Generation Computer Assisted Prewriting. In *The Computer in Composition Instruction*, W. Wresch (Ed.) IL: National Council of Teachers of English, pp. 15–33.
Campbell, J. 1988. *The power of myth*. NY: Doubleday.
Cassell, J. 1998. Storytelling as the Nexus of Change in the Relationship between Gender and Technology. In J. Cassell and H. Jenkins (Eds.), *From Barbie to Mortal Kombat: Gender and Computer Games*. Cambridge, MA: MIT Press.
Cassell, J. 1995. The role of Gestures in stories as multiple participant frameworks. *AAAI Spring Symposium: Interactive Story Systems*, pp. 23–26.
Colby, K. 1975. *Artificial Paranoia: A computer Simulation of Paranoid Processes*. NY: Pergamon Press.
Daiute, C. 1985. *Writing and Computers*. Reading, MA: Addison-Wesley.
Davidson & Associates. 1995. KidWorks Deluxe. CD-ROM for Windows/Macintosh.
Domeshek, E. 1992. *Do the right thing: a component theory for indexing stories as social advice*. PHD Thesis. Northwestern University.
Don, A. 1990. Narrative and the interface. *The art of human-computer interface design*. NY Addison Wesley.
Druin, A. 1987. *Building an alternative to the traditional computer terminal*. SM Thesis. MIT, Cambridge, MA.
Dyl, J. & S. Wapner. 1996. Age and Gender Differences in the Nature, Meaning, and Function of Cherished Possessions for Children and Adolescents. *Journal of Experimental Psychology*, 62, pp. 340–377.
Erikson, E. H. 1950. *Childhood and Society*. NY: Norton and company.
Freud, A. 1965. *Normality and Pathology in Childhood*. London: Hogarth Press and the Institute of Psycho-Analysis.
Foner, L. 1998. Are We Having Fun Yet? Using Social Agents in Social Domains (This volume).
Fuller, M. & H. Jenkins. 1995. Nintendo and New World Travel Writing: A Dialogue. G. Jones (Ed.) In *Cybersociety: Computer-Mediated Communication and Community*.
Glos, J. 1997. *Digital Augmentation of Keepsake Objects: A place for Interaction of Memory, Story and Self*. SM Thesis. MIT, Cambridge, MA.
Gombert, J. E. 1992. *Metalinguistic development*. Chicago: Univ. of Chicago Press
Hickman, M. 1987. *Social and Functional Approaches to Language and Thought*. NY: Academic Press.
Harel, I. & S. Papert (Ed.). 1993. *Constructionism*. NJ:Ablex Publishing.
Ishii, I. & B. Ullmer. 1997. Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *CHI'97 Proceedings*, ACM, pp. 234–241.
Karmiloff-Smith, A. 1985. Language and Cognitive Processes from a Developmental Perspective. *Language and Cognitive Processes*, 1, no. 1, pp. 61–85.
Laurel, B. 1993. *Computers as Theater*. NY: Addison Wesley.
Linde, C. 1993. *Life Stories. The creation of coherence*. NY: Oxford University Press.
Mauldin, M. 1994. Chatterbots, TinyMuds & Turing Test. 12th Conference on AI.
Miller, G., R. Beckwith, C. Fellbaum, D. Gross, & K. Miller. 1993. Introduction to WordNet. http://www.cogsci.princeton.edu/uwn.

Miller, P., R. Potts, H. Fung, L. Hoogstra, & J. Mintz. 1990. Narrative practices and the social construction of self in childhood. *American Ethnologist*:17(2) pp. 292–311.

Montford, N. 1998. *A conversational Computer Character to Help Children Write Stories*. SM Thesis. MIT, Cambridge, MA.

Murray, J. 1991. Anatomy of a New Medium: Literary and Pedagogic Uses of Advanced Linguistic Computer Structures. *Computers and the Humanities*: 25, pp. 1–14. Netherlands: Kluwer Academic Publishers.

Papert, S. 1980. *Mindstorms: Children, Computers and Powerful Ideas*. NY: Basic Books.

Papert, S. 1987. Computer criticism vs. technocentric thinking. *Educational Researcher*: 16 (1), pp. 22–30.

Papert, S. & M. Resnick. 1995. Technological Fluency and the Representation of Knowledge. Proposal to the National Science Foundation. MIT Media Laboratory. Cambridge, MA.

Polanyi, L. 1989. *Telling the American Story*. Cambridge: MIT Press.

Polkinghorne, D. 1988. *Narrative knowing and the Human Sciences*. State University of NY Press.

Resnick, M, A. Bruckman, & F. Martin. 1996. Pianos not stereos: creating computational construction kits. *Interactions*, Sept.-Oct. 1996, pp. 41–49.

Rosen, S. 1982. *My voice will go with you: The teaching tales of Milton H. Erickson*. NY: Norton.

Schank, R. & C. Riesbeck. 1981. *Inside Computer Understanding: Five Programs plus Miniatures*. NJ: Lawrence Erlbaum.

Schank, R. & R. Abelson. 1995. Knowledge and Memory: The Real Story. *Advances in Social Cognition*, Volume VIII, pp. 1–85. NJ: Lawrence Erlbaum.

Suchman, L. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. NY: Cambridge University Press.

Turkle, S. 1984. *The Second Self: Computers and the Human Spirit*. NY: Basic Books.

Turkle, S. 1995. *Life on the screen: Identity in the Age of the Internet*. NY: Simon & Schuster.

Turner, V. 1980. Social Dramas and Stories about Them. In W.J.T. Mitchell (ed) *On Narrative*. Chicago: University of Chicago Press.

Umaschi, M. 1996. SAGE Storytellers: Learning about Identity, Language and Technology. In *ICLS 96 Proceedings*, AACE, pp. 526–531.

Umaschi, M. 1997a. Soft Toys with Computer Hearts: Building Personal Storytelling Environments. In *CHI'97 Proceedings*, ACM, 20–21.

Umaschi, M. 1997b. *Soft interfaces for interactive storytelling: learning about identity and communication*. SM Thesis. MIT, Cambridge, MA.

Weizenbaum, J. 1976. *Computer power and human reason*. SF: Freeman & Cia.

Wigren, J. 1994. Narrative Completion in the Treatment of Trauma. *Psychotherapy*, vol.31, no.3.

Winnicott, D. W. 1971. *Playing and Reality*. London: Tavistock Publications.

Winograd, T & F. Flores. 1986. *Understanding Computers and Cognition*. NY: Addison Wesley.

CHAPTER FOUR

Autonomous Synthetic Computer Characters as Personal Representatives

Linda Cook, Tim Bickmore, Sara Bly, Elizabeth Churchill,
Scott Prevost, and Joseph W. Sullivan

1. Introduction

The expression "you can't be everywhere at once" aptly describes the motivation underlying the various artifacts that individuals have created over the years to represent themselves in their absence. It also describes our interest in developing novel forms of personal representation that *will* enable people to be everywhere at once (or almost everywhere). Familiar representational artifacts developed over the years to convey a pseudo-physical presence include resumes, application forms, portfolios, answering machines, email, and web pages. The research described in this chapter focuses on the design and development of autonomous synthetic computer characters to serve as a more robust form of personal representation. Our ultimate expectations for these synthetic characters far exceed what other representational artifacts have been able to achieve thus far. We believe that in the near future one's 'embodied' personal representative will be able to:

- Present the likeness of the author (both visually and behaviorally) to others
- Represent the author's interests to others and to act on their behalf
- Increase an author's sense of presence by using the metaphor of face-to-face communication when interacting with others
- Be autonomous or capable of operating without the direct control of their

authors, operating opportunistically and reactively, and having their own goals and motivations (as defined by their authors)
– Be easily authored by users who have little or no programming experience

This chapter begins with a brief historical review of the directions that our research has taken over the past two years. We discuss some of the different application domains for which autonomous synthetic characters were considered and the research literature that helped frame our research focus and provide direction for applications development. Section 3 provides a more detailed discussion of the role that self representation currently plays in internet related forums like MUDs, MOOs, web pages and computer mediated technologies that serve groups who collaborate across distance and time. Section 4 discusses basic design considerations in creating autonomous synthetic characters such as visual appearance, autonomy, psycho-social issues, multi-modal communication and authorability. The chapter concludes with a description of document avatars — synthetic computer characters that can be attached to documents to express the personal views of document's authors.

2. Research Goals and Application Domains: A Brief History

One goal of this research project was to create an end-user authoring system that would allow users to create a synthetic computer character capable of acting on their behalf in some limited capacity. We envisioned the characters as semi-autonomous in their behavior, having a personality and a visual likeness similar to their author, social intelligence and, finally, the capacity for some degree of multi-modal communication. The platform for our initial prototype applications was the internet and, in particular, personal web pages because they were a technology-rich medium for self-representation. Web pages have global reach, the ability to interact with multiple users at one time, support a number of different communication modalities, and require minimal, low cost effort to develop and publish. In addition, most personal web pages today are probably underutilized — consisting primarily of static information (and images) and possibly a "mail-to" link for messaging the web page author. While the current state of web page technology represent a tremendous improvement over other forms of representation their potential for more versatile and compelling forms of self representation remains largely untapped.

A second goal was to gain insight into the construct of self representation

and what it means to create a virtual version of oneself. This is (and will continue to be) an issue that will be far more complex than the technical difficulties likely to be encountered in this project. There are volumes of academic work devoted to this topic and numerous self-help texts for gaining insight into one's psychological makeup, recreating one's personal image or both. Even the simple representational artifact, the resume, can create great discomfort for individuals as they try to anticipate how they will be perceived (and judged) by intended and unintended features in the presentation of themselves. Clearly, generating a character-based self representation will be subject to similar, but, far more complex social and psychological criteria.

The next two sections present a brief history of this research over the last two years. The first section examines some of the application domains for personal representatives that were initially considered. These application prototypes (in various stages of analysis and development) were a way to evaluate some of the affordances synthetic computer characters might have in different domains. The second section discusses the various research disciplines that contributed to our understanding of the task of making our characters not only personable but also believably compelling.

2.1 *Selected application domains*

To enable users to create their own personal representative in the form of a synthetic character required a character authoring tool capable of the following: the ability to create a physical rendering of the character, an animation system or scripting language to enable the character to behave in specific ways and a system to create the physical environment in which the character could live. We have focused on applications that could be used to gage the acceptance and value of animated autonomous personal representatives in the workplace. One of the first prototype applications considered was a *cyberbiography*. Its purpose was to enable an *author* to capture personal knowledge and affect (behavior and emotional state) in a web-like document. *Readers* could review and read the cyberbiography with the aid of the author's synthetic character. The authoring environment for the end-user consisted of two components: a character design studio and a scene design studio.

Each cyberbiography was made up of a series of *scenes* (analogous to web pages) that encapsulated information for presentation. Scenes consisted of three components: props, affects and behaviors, and queries. *Props* were

objects which could exist in a character's environment and with which the character could interact. Typical examples of props were: documents such as white boards, easels, pictures and video clips; environmental backdrops such as office and conference rooms; and character-specific props like a desk, chair or other personalized items. *Affects and behaviors* were explicit, grossly animated behaviors that could be scripted by users. Behavior sets were intended to not only provide basic animation capability but to also convey affective state and emotions on the part of the character. Some generic classes of behavior included:

- speak directives – pre-recorded audio clips by the author to maintain of the author's sense of presence
- refer-to-directives – to direct readers attention to a prop (i.e. author's character points to an object),
- emotion directives – character could display emotional state (i.e. happy, sad)
- animation directives – low level-animation sequences such as walk-to

One simple aspect of social intelligence that we examined was enabling authors to create a character which could present a slightly different self-image depending upon the social relationship of the author to a reader (i.e. colleague, boss, visitor). For example, most authors would probably choose to present a different image to a co-worker than to their boss. Readers were classified on the basis of their email address.

To enable the synthetic character to communicate with readers, *queries* were built into a menuing system. These were a set of questions the reader could ask (i.e. Where is Tim?) or a set of actions that the reader might ask the character to take within a given scene (i.e. Reserve a meeting time with Tim).

The elements found in the cyberbiography have largely been carried over to a similar application prototype called document avatars. These are user-authored, synthetic computer characters that are used to annotate documents with an author's unique point of view. In this system users can author their own personal representatives and attach them to documents in order to express their personal views. Document avatars (Bickmore, Cook and Churchill 1998) can be interacted with by readers of the document in several ways to obtain the opinions of or explanations about parts of the document, or to get guided tours through the document. The system also allows multiple representatives to be attached to a single document to provide the points of view of multiple authors (e.g., in a collaborative effort). As one of the most fully developed systems it is

discussed in more detail in Section 5.

Our current application prototype is Kumo — an assistant for users of an advanced, high-tech audio visual system in a conference room at FX/PAL. Kumo will provide users with tutoring or general assistance with highly sophisticated audio-visual equipment. Our research is largely focused on issues involved in embodying a character with human-like qualities to serve as an anthropomorphic, face-to-face communicative interface. Some of Kumo's abilities will include: face-to-face communication, autonomous behavior, multi-modal communicative functions, natural language understanding and generation, gesture recognition and individuated personalities.

2.2 *Research focus*

Our approach to the psycho-social aspects of personal representation draws from two research areas: psychology (specifically theories of personality, communication theories and social cognition) and computer human interaction. Current models of personality in autonomous characters are able to convey somewhat realistic, but limited and stereotypical behaviors. The behaviors of most characters are generally scripted (i.e. little autonomy) and are not authorable. Our model of personality attempts to capture personality expression as a form of character individuation — characters take on (and sustain) unique and identifiable personalities through behaviors that are consistent with a given personality profile. The model draws from personality trait theories and nonverbal communication theories which are embedded in a social-cognitive model of personality.

Our analysis of the literature on computer human interaction began by focusing on the role personal representation plays in internet-based technologies such as web pages or virtual environments. When given the choice, do people naturally create personalized images to convey a sense of their unique identity and, if so, in what ways are those images created, displayed and received by others. More recently, we have focused on the field of computer mediated collaborative work environments — the design of tools to facilitate more 'face-to-face communication' opportunities for individuals who are not collocated and often unable to communicate synchronously (i.e. in real time). This field is particularly important to our efforts as it represents groups of individuals who are, in varying degrees, unable to be physically present during collaborative work efforts. Our interest in this field up to this point has

been primarily oriented towards data collection — in particular the construction and use of personal web pages in work settings. However, the large body of research in this field and the applications that have been developed represent a wealth of ancillary information that can be important in the future development of synthetic characters as personal representatives. Section 3 covers some of the issues associated with self representation in different forums on the internet and in computer-mediated work environments.

3. The Role of Self Representation and Computer-Mediated Technologies

We need look no further than the way individuals use an email facility to conclude that there is an interest on the part of people to convey a personal identity. Some email users are satisfied with an 'image' or a personal identity that is nothing more than the sender information that normally appears in their message header. Others, however, create specially formatted return addresses or select quotes to convey their own unique identity. Work by Donath (Donath 1999) provides a detailed analysis of the anatomy of Usenet letters (account name, voice and language of letter content and the signature line) and how these elements do provide identity anchors from the 'virtual persona to the real world person', whether intended or not. In other internet forums users are more consciously aware of and actively involved in the construction of special online identities; assuming a 'physical identity' in virtual environments. As is described in more detail below, work-related web pages almost always display some form of personal representation that uniquely identifies its author.

Our discussion begins with a brief review of personal representation in MUDs (Multi User Domains) and MOOs (MUDs, Object-Oriented) — a forum that closely parallels our interests in developing embodied personal representatives. Following the discussion on virtual environments is a selected review of research from computer-mediated collaborative work spaces. Topics include the effects of distance on the perception of others and the problems of collaboration in work environments in which individuals are distributed and communicate asynchronously (i.e. not in real-time). Environments such as these most closely approximate working conditions in which one is unable to be physically present and, thus, might benefit from a 'sense of presence' conveyed by a synthetic character acting as a personal representative.

3.1 *MUDs and MOOs*

MUDs and MOOs are virtual environments in which characters, created by users, can interact with each other in a manner that can be realistically compelling. These environments illustrate some of the best forms of self-representation that have been consciously constructed by users. For example, in online virtual worlds such as OnLive and the Palace (OnLive!; The Palace) users can represent themselves through graphical embodiments or 'avatars', which are essentially remote-controlled digital puppets. In these systems users can choose a representation by constructing customized avatars, or selecting them from libraries, and then animate their behavior to varying degrees.

One of the most informative studies conducted on issues related to personal representation in MOOs was done by Shiano and White (Schiano and White 1998). Their research examined active participants of LambdaMOO, a social MUD of about 7000 members, using online surveys, personal interviews and logging studies (1 minute intervals, 24 hours/day for two weeks). Of particular interest was the data that dealt with issues of identity: visual appearance, identity and role-playing and user motivation for participation in MUD.

– *Visual Appearance.* In the initial stages of our research we assumed that choosing an animated character as a form of self representation would be interesting to some of the more adventurous and perhaps growing more slowly in popularity as synthetic characters became more commonplace. In this study people expressed more interest than expected in animated forms of representation. For example, when given the choice of text, audio, video or animation as forms of representation the first two choices were text followed by animation. (Women tended to choose text, men were equally split between text and animation.) Surprisingly, users also indicated a preference that the character look like them; with women preferring a realistic portrayal and men preferring image-based. This is in contradiction to the generally held assumption that an attractive feature of this type of forum is the anonymity afforded by on-line characters (Turkle 1995).

– *Identity and Role-Play.* People tend not to use multiple characters as had been commonly thought. Only half had more than one character and those with more than one averaged only two. (Secondary morphs tended to be short-term, costume-like or comedic in nature rather than alternate identities.) Also interesting was that most users indicated they 'were themselves', or a slightly

idealized version of themselves, and were not role playing. The maintenance of a personalized identity was fairly consistent and might be attributed to the fact that users revealed some social pressure to maintain a single identity as a form of accountability to the user community as a whole. As suggested by the authors, this finding is remarkably similar to Goffman's contention that 'long-term social cohesion requires stability of self-presentation' (Goffman 1995).
– *The Social Nature of a MOO*. The primary reasons for MUDing were social in nature — probably not an unexpected finding. However, in light of the tendency to prefer a form of personalized identity their socialization efforts were largely as themselves and probably not unlike that in their real world environments. Socializing was generally in small groups and privately conducted. Almost all users augmented a sense of presence in the community by having private places (or "homes"); spending about an equal amount of time there or in public rooms. Personal interactions were conducted in private circumstances.

3.2 *CSCW — Computer supported collaborative work*

Computer mediated collaborative work environments are designed to replace or complement face-to-face communication for individuals who are not physically located together (i.e. distributed) and often unable to communicate synchronously (i.e. in real time). Both research and application development in this field focus on creating computer-mediated technologies that enable non-collocated individuals to work in a manner that more closely approximates the rich environment afforded by collocation and face-to-face communication. Most recent applications (called GroupWare) use the real-time connectivity of the internet to provide workers with a broader communication bandwidth to overcome the limitations caused by their physical separation. Some application prototypes include communication in real-time via on-line chat facilities and the simultaneous sharing of common work objects and artifacts such as design schemes, structured drawings or graphics (Stewart, Raybourn, Bederson and Druin 1998; Vertegaal, Harro and Slagter 1998; Greenberg 1997).

The motivations underlying our interest in this area were both the research findings that might be relevant to better understanding issues of personal representation and a more long-range goal of identifying potential applications in which autonomous synthetic characters might be deployed. Some relevant research findings include such things as how distance and an

identity of a text string email address affect peoples' behavior in the work place, their initial perceptions of others with whom they might work or the social cohesion that may arise among potential group members. Finally, one area in which there may be many potential applications for autonomous characters is the increasing number of distributed groups who are limited to communicating asynchronously. In these environments the real-time connectivity afforded by the internet (i.e. chat facilities, shared artifacts) is of little use (i.e. 10 a.m. on the west coast of the United States is 2 a.m. in Japan).

3.2.1 *Distance, anonymity and the perception of others*

Personal interactions within a work environment are important for establishing trust, facilitating cooperation and for exploiting individual knowledge (Bly, Cook, Bickmore, Churchill and Sullivan 1998). Likewise, a lack of personal interaction within a work environment can present significant obstacles. Research described in more detail below shows that lack of face-to-face communication can affect a group's underlying social relationships, cooperation, honesty and openness to input from more distant participants. Our own research on aspects of personal representation in individual web pages at work suggests that, even though web pages are used to organize and disseminate project-related information, people actually prefer pages with a more personal orientation. We review three studies on trust, distance, web pages at work.

– *Trust.* Both experimental and informal studies show that people are less willing to use GroupWare for tasks which require mutual trust among group members (Rocco 1998). The general assumption is that lack of face-to-face contact among members reduces their trust of others and their commitment to the group process. In this study three groups of subjects (all strangers) participated (via synchronous email) in a social dilemma task for which there was a strong incentive for individuals to defect for personal gain rather than cooperate for the group's good. One group had a face-to-face discussion about the experimental task for 10 minutes prior to participating in the task itself, a second group participated face-to-face in an in-group socialization task (not related to the social dilemma task) one day prior to the experimental task and, a final group was allowed no face-to-face contact prior to the experimental task (i.e. task was performed entirely by email communication). Both groups with face-to-face communication reached a cooperative consensus between the 10th and the 13th trial while the email only group never achieved a

successful level of group cooperation.

Two aspects of her findings were particularly important to our interests. First, the relatively successful performance of the group which met face-to-face for the in-group socialization task suggests that the presence of a social network between group members greatly improves overall cooperation and, thus, performance. This supports suggestions made by (Nohria and Eccles 1992) that the effectiveness of GroupWare technologies (in this case email) may depend to a large extent upon establishing some underlying network of social relationships. A second interesting result was the communication patterns in the email only group. Unlike the other groups, conversational coherence in the email-only group was absent, resulting in multiple parallel threads of discussion attributed largely to a lack of sequential turn taking. Rocco suggests that email-only subjects could avoid answering questions and ignore messages because the sender lacked a 'physical presence' which in a real world setting would generally compel a response (i.e. senders were essentially anonymous to the receivers).

Under normal circumstances (i.e. collocation) social networks and a sense of physical presence would be a natural by-product of the day-to-day communication and interactions that occur in such settings. With distributed groups communicating asynchronously it is less clear how such social networks can develop. One possibility is more personalized access to other group members such that individuals begin to take on their own unique and sustainable identities that become familiar to others in the group. This might mitigate some of the negative aspects of distributed groups described by Rocco: depersonalization, lack of contextual presence and the absence of non-verbal cues that signal willingness for teamwork and good will.

Distance: Deception and Persuasiveness. Work by (Moon 1998) found that as the distance between an individual and a computer mediated interviewer increased (local, San Jose, Boston) so did the inclination of the individual to give inaccurate responses that over-reported desirable behaviors and under-reported undesirable behaviors. In short, the greater the perceived distance, the greater the tendency to distort responses so as to present oneself in a more positive light. Not only does perceived distance seem to affect what individuals may project about themselves, it may also affect how receptive they may be to accepting input from others. Individuals were more likely to accept suggestions and advice from a computer source seen as physically proximate than by one who was perceived to be more distant. This finding resonates well

with that of Rocco — people give less credence to those for whom they have a remote sense of their physical being (i.e. not responding to email messages in the email only group). Both studies are also similar to the general findings that anonymity may actually encourage non-cooperation because of a lack of individual accountability (Wiggins, Wiggins and Vander Zanden 1994).

Web Pages at Work: Increasing Inter-Personal Interactions. Research by our group, (Bly, et. al. 1998), found that personalized web pages as a form of mediated communication within work settings might promote more *inter-personal* interactions for individuals. She examined the web pages of employees in three high-tech product companies and one college looking at the ability of web pages to capture and present the personal expressions of their authors in conjunction with project or task-oriented information. That study revealed interesting aspects about personalization that were important not only for authors, but also when authors became 'readers' of web pages.

Not surprisingly, personal web pages were most often used in support of an employee's project work: information dissemination, project tracking, and information management for the individual him/herself. However, authors almost always included personal information and pictures on their pages (to "provide a sense of themselves"). Even images other than photos were clues to the person (i.e. one author involved in testing had a graphic of a bug on his page). As readers, most interviewees found value in the personal information about the authors of web pages — whether or not they were willing to provide personal information themselves. Some typical comments included "a personalized page lets you get to know someone a little better" or "I get more of a person if I can see the face".

The relevance of research like that described above to embodied computer characters as personal representatives is only conjecture at this point but it does highlight some directions for future research on their potential usefulness. For example, would there have been a geographical difference between computer interviewers (local, San Jose or Boston) if the interview had been conducted by an autonomous personal representative (from each location) appearing *within* the subject's interface? Likewise, if strangers could interact briefly with the personal representatives of others in the group and maintain visual contact with them during task performance would they be more responsive to other member's email or increase their cooperation with others in the group. Finally, people actually seem to like the more personal aspects of their colleagues' web pages and use those in ways they believe lets them get to

know the authors a little bit. If a single picture can convey a sense of presence would not an embodied character add a significant degree of realism to an absent author's electronic presence?

3.2.2 *Asynchronous, distributed communication and shared artifacts*

Numerous articles have pointed out the costs to collaboration when groups are separated simply by distance: gestures that express ideas, signal turn-taking or focus the attention of the rest of the group (i.e. pointing); contextual sharing between group members; shared artifacts as a central focal point (Olson, et. al 1998); gaze awareness (ability to monitor direction of group member gaze); awareness of others and so on. The absence of this broad bandwidth of real-time communication is the motivation underlying the development of various communication widgets designed to support electronically sharable artifacts, telepresence, and real-time chat facilities. Unfortunately, this technology is less applicable for those environments in which group members are separated by both distance and time. This environment is of most interest to our research because it best illustrates the conditions under which an autonomous character acting as a personal representative may be most valuable.

Real-time collaboration across distance and time will be feasible when autonomous characters reach a level of sophistication that enables them to represent their absent authors during computer-mediated concurrent collaboration. The level of character sophistication required for this is in the distant future, but, in the interim, synthetic characters are capable of performing some of the real-time communicative functions that may be absent in these environments. For example, synthetic characters created for our research purposes are capable of:

- Verbally expressing an author's point of view which, in conjunction with appropriate nonverbal behaviors, can emphasize the most important aspects of the author's message.
- Displaying a wide range of gestures towards shared artifacts such as pointing or making group members aware of the author's gaze to suggest current focus of attention.
- Serving as a type of daemon for artifact modifications such that when an author's strongly held views are challenged by a change in the state of the artifact the views of the author can be made known. (Information in this case is made available only when necessary.)
- Being queried by other group members about specific views that might be held by the author so as to explain, recommend or justify the author's

position.
- Increasing the amount of shared context among group members — since multiple characters can be attached to a single artifact it provides each member with concurrent-like input from others in the group.
- Increasing the sense of awareness of others (i.e. a sense of presence), giving individuals a greater sense of legitimacy within the group.

A subset of these capabilities are a part of the document avatars discussed in more detail in Section 5 and in (Bickmore, Cook and Churchill 1998).

4. Basic Design Considerations in Autonomous Synthetic Computer Characters

The construction of autonomous personal representatives represents a vast research area touching on issues in four major theoretical areas: visual representation (synthetic characters), multi-modal communication (including both verbal and non-verbal behaviors), behavioral representation (system autonomy), and the psychological issues in creating an electronic 'self' (self representation). Also crucially important, however, is how synthetic characters can be made authorable by end users. In this section we outline some of the key research issues in each of these areas.

4.1 *Synthetic characters*

Synthetic characters are graphical representations of humanoids or other creatures in 2D or 3D and are animated. Research has focused not only on how to make the appearance of these characters more lifelike, but also on how they can be given the ability to operate autonomously (Blumberg and Galyean 1995; Blumberg 1996). The issue of a character's autonomous behavior is of fundamental interest to our research. As described previously, the use of visual representations of oneself within a virtual worlds environments on the internet, such as OnLive and the Palace, has attracted a large number of users. Figure 1 illustrates several custom designed avatars in the Palace in use at FX/PAL. However, self representation in these systems is through avatars which are remote-controlled digital puppets. As such, they provide no useful functionality once the puppeteer walks away from the keyboard. This issue of autonomy is discussed in more detail below.

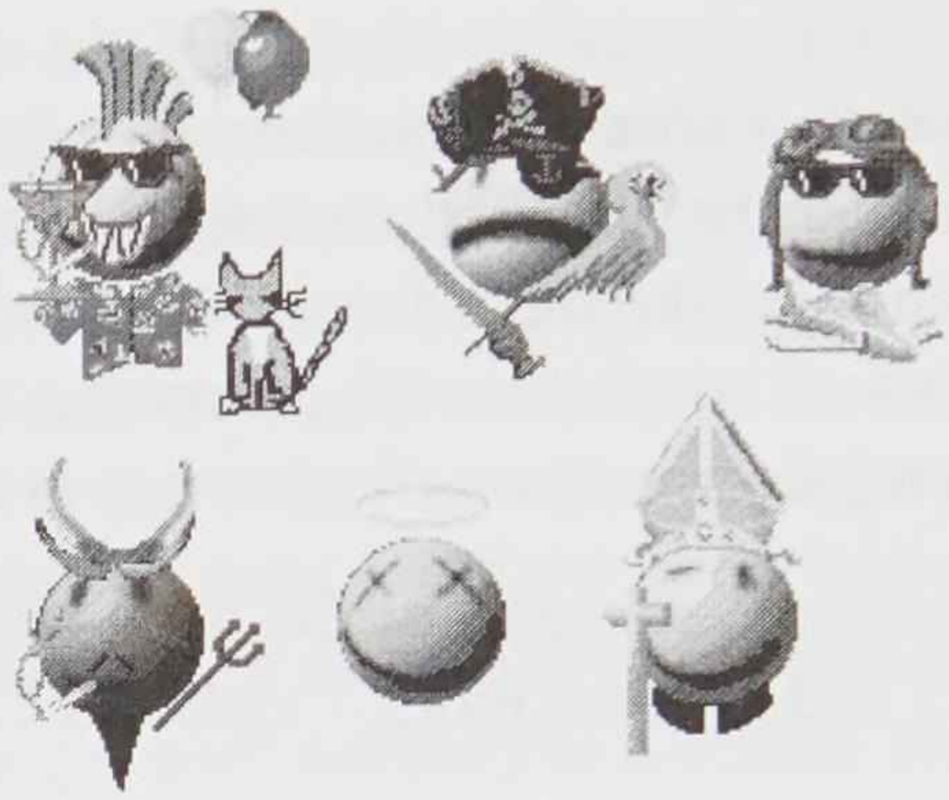


Figure 1. Examples of Avatar Individuation in the Palace

What should (or will) determine a character's physical appearance is hard to predict. However, as described previously, evidence suggests that most users prefer a representation they can identify with (Schiano and White 1998; Worlds Chat). Other research suggests that, when given a choice, users want a personal representative with their visual likeness as most personal web pages contain photos of their authors (Bly, et al. 1998).[1] Because we are primarily interested in applications in professional work environments we assume that end users will probably want realistic representatives. However, the model from virtual worlds of changing one's likeness (Stone, Fishkin and Bier 1994) could be applied to personal representatives as well. For example, an end user may specify multiple representations for themselves which iconically encapsulate the particular representative's expertise (e.g., "Carleen the Administrator" vs. "Carleen the Recreation Coordinator").

4.2 *Autonomous systems*

What does it mean for one's representative to be autonomous? Several criteria are specified in (Maes 1994), the most relevant of which are that the representative must be able to operate independently from its author, and must be able to fulfill its goals in a complex dynamic environment. The range of behaviors that an autonomous representative could have include:

- Providing information about its author;
- Voicing the author's opinions and personal preferences;
- Relating and applying the author's expertise in certain narrow domains;
- Relaying messages to the author; and
- Making commitments on behalf of the author.

Considering the use of autonomous personal representatives in general, there are many interesting research questions and issues to be addressed. For example, how much authority should a representative be granted? Can it make commitments on behalf of its author? Should it be trusted with a credit card? How good is a representative's word?

If an author's representatives are truly *adaptive* autonomous agents (Maes 1994), is it important that the things they learn be shared with each other to ensure consistent behavior? Should this information always be fed back to the author as well? And if a user interacting with a representative knows that its author is watching the interaction (synchronously or asynchronously), how does that change the nature of the interaction — such as the notion of de-individuation in psychology (Prentice-Dunn and Rogers 1982)?

While our current implementation of personal avatars utilizes purely scripted behavior, we have started development on a framework which will allow truly autonomous behavior while still providing for various degrees of end-user authorability and direction. The framework is based on an architecture for conversational characters that we have developed for a variety of applications (Cassell, et al 1998; Churchill, et al. 1998). The architecture includes modules for both reactive, real-time behavior-based control of a synthetic character, in addition to non-real-time deliberative processes, including task and communication planning and multi-modal input understanding and generation. The reactive part of the system is inspired by the ALIVE architecture [Blumberg,1996] which selects a behavior for the character to perform on each update cycle based on internal state (endogenous) variables, external stimuli (releasing mechanisms), past behavior, and influences from any currently executing plan. Behaviors are arranged in an abstraction hierarchy, with behaviors at each level competing for control of the character. The reactive module can pass goals to the planner to be achieved (in addition to persistent goals), and generated plans serve to strongly influence the selection of behaviors in the reactive module. In addition, plan and plan step preconditions and constraints are monitored by the planner so that obsolete plans can be discarded and re-generated as necessary.

4.3 *Multi-Modal Communication*

Anthropomorphic interfaces provide the opportunity for personal representatives to use the metaphor of human face-to-face communication in user

interactions. Given the appropriate sensory inputs, such representatives could engage in multi-modal discourse with users approximating face-to-face conversation with the author. This would include the recognition and use of gestures, facial expressions and body posture, and the more traditional verbal channels afforded by speech and natural language understanding and synthesis modules (Thorisson 1996b).

Some of the unique research issues in this area relative to personal representatives are:

- How to trade off the personalization afforded by recorded speech with the flexibility of synthesized speech;
- At what level should a representative's communicative acts be scripted vs. being generated from more general communicative goals;
- Which aspects of communicative behavior are generic and which can be tailored to achieve personalization?

4.4 *The social and psychological aspects of self representation*

We took a very pragmatic approach to the role that psychological issues would play in the development of synthetic characters at FX/PAL attempting to ensure that the synthetic characters we built for different applications were 'socially intelligent'. Initially, our common sense definition of a socially intelligent agent was one that did not offend and had enough rudimentary social skills to be taken seriously by those with whom it interacted. After much research and some early development effort in this area social intelligence came to mean many different and important aspects of character development and personification: personality, behavioral manifestations of personality, social skills, impression formation (i.e. as others might see a character), impression management (how the character desires to be perceived), issues of individual differences and so on. To gain a sense of coherence over the broad ranging and diverse components that comprise social intelligence we relied on a social-cognitive model of personality proposed by Hampson (Hampson 1988; 1992). Its theoretical approach enabled us to unify a subset of components important in building believable characters and also provided guidance during preliminary design and implementation efforts. Ultimately, we hope the model will lead to the design of more wholly integrated and compelling characters.

Hampson's approach views personality from both *within* an individual (i.e. an internal disposition that guides decision making, biases perceptions)

and *between* an individual and those with whom the individual interacts. In other words people ascribe personality attributes to others based on the perception and interpretation of an individual's behavior. His model has three interrelated components: actor (an individual's internal characteristics and dispositions), observer (the interpretation of the meaning of an actor's behavior by *another person*) and self-observer (an actor attempts to identify and manage others' judgements). This model captures a number of common sense ideas about personality — it is largely static and unchanging, people can readily label and classify others' personality based on behavior, and people can modulate different aspects of their personality.

Using this model as a guide we looked for disciplines that addressed the first two components (actor and observer). The theoretical basis for modeling the actor was derived from trait theories of personality. This approach suggests that personality can be described by a small set of psychological traits that are well known and understood by most people. Each trait represents a behavioral dimension that can be quantified along a continuous scale like introvert/extravert or warm/cold and a set of behaviors that describe each end of the continuum. For example, most people would describe introverts as being quiet, shy and retiring. Differences in personality are reflected by the fact that people tend to vary widely along the continuum within each trait. Thus, for example the introverted person who is also warm has a different trait profile than an individual who is also introverted but seen as cold.

However the 'label' of introvert does not suggest exactly 'how' a character must behave to be labeled as an introvert. The field of human communication is the study of how humans both emit (or encode) personality attributes and as viewers decode personality attributes from behavior. As observers, people read from and combine information across multiple nonverbal behavioral channels (i.e. facial expression, body posture, gestures) to form an overall impression of another person (Lathers 1997). Thus, research has shown that introverts tend to display the following kinds of behavior: they speak little and only when spoken to, they assume a closed body posture and they tend to avoid eye contact. If a person (or an embodied character) were to display these behaviors they would be labeled as introverts. Our methodological approach is to isolate a basic set of personality traits and to build a database of both nonverbal and verbal behaviors that are 'trait-indicative' (i.e. people believe the behavior reflects a specific trait). Personality can then be displayed by sampling from those behavior sets that predominantly reflect a character's personality trait profile.

4.5 *End user authoring*

End-user authoring, while tangential to the usage and architectural issues of autonomous personal representatives, is crucially important if we expect these systems to be used by anyone except for programmers. Although some research on end-user authoring of autonomous systems has been carried out (Blumberg and Galyean 1995; Blumberg 1996) there is little research regarding the unique authoring goals that we are confronted with here: enabling an author to not only guide the sequence of behaviors needed to convey their intended message but also to deliver that message in much the same manner as the they would in real life.

Specifically, we are concerned with two levels of behavioral authoring. The first level includes basic motor behaviors (i.e., moving from point X to point Y), gestures and facial expressions. These activities are general to all users. The second level of behavioral authoring clearly requires finer grained design decisions on the part of end users and therefore is a more difficult problem with respect to the authoring tools required. This level involves the individuation of personal representatives, that is, the ability to express the subtle individual differences that uniquely define the author. Ideally, an identical sequence of behaviors performed by two different personal representatives should reflect the individual differences found in their authors. For example, a simple walk across the screen by an individual who is more formal and dominant will be vastly different than a walk by a individual who has a more casual and relaxed attitude.

5. Application: Document Avatars

To gage the acceptance of animated autonomous personal representatives in the workplace, we have constructed a prototype system in which users can easily author representatives and attach them to documents in order to express their personal opinions about the document's contents. One of the motivations for this particular application is the fact that an individual can often present an idea in a 10–minute personal presentation that would otherwise take pages of formal documentation to describe. Also, attaching a representative to a document restricts its domain of discourse to the subject matter of the document, providing a context in which it can exhibit complex behavior while still being authorable in a short period of time.

Document Avatars can be interacted with by readers of the document in several ways to obtain the opinions of or explanations about parts of the document, or to get guided tours through the document. Other research projects which have investigated automated presentation of document contents include the PPP Persona project (Andre, Muller and Rist 1996) and Jack Presenter (Noma and Badler 1997) neither of which addressed character individuation or authorability, and the Active Paths system (Zellweger 1988) which did not utilize a synthetic character.

The system also allows multiple representatives to be attached to a single document to provide the points of view of multiple authors (e.g., in a collaborative effort). This provides functionality which is similar to that provided by the Magic Lens system (Stone, Fishkin and Bier 1994), in that multiple views onto a single document are supported.

Although the behavior of these representatives is currently scripted, we feel that we can learn important lessons from the feedback we get in using this system to inform our design of an action selection engine which produces truly autonomous behavior while providing the proper level of end-user authorability.

5.1 *Representation*

Document avatars are created in three parts; their visual appearance, their behavior, and recorded audio files. We felt that recorded audio was superior to synthesized speech both in overall quality, and that it aided in the personalization of the representative avatars.

The system uses 2D rendering for avatar animation using warp-able polygons (Thorisson 1996a). In this approach, an avatar consists of an arbitrary number of polygons of arbitrary complexity, color, fill, and layering. Each polygon can be attached to one or more control points, which change the shape of the attached polygons when moved during rendering. A "pose" consists of a named set of locations for some number of control points. Run-time animation is achieved primarily by interpolating between poses over specified time intervals. Several poses can be executed simultaneously, giving an avatar the ability to "smile," "point," and "walk" at the same time, for example. This approach to animation was selected because it simplifies end-user authoring and provides a run-time animation engine which is computationally efficient.

The visual appearance of an avatar is created in the Avatar Studio, shown in Figure 2. This allows users to construct polygon-based avatars in a "MacDraw"-like environment. The novel feature of this tool is the ability for authors to import a digital photograph of themselves, which is made semi-transparent and overlaid onto the drawing area. This allows them to simply trace polygons over the image to very quickly and easily construct their avatar. The output of the Studio is a compact text representation of the polygons, control points, and poses which comprise the visual appearance of an avatar.

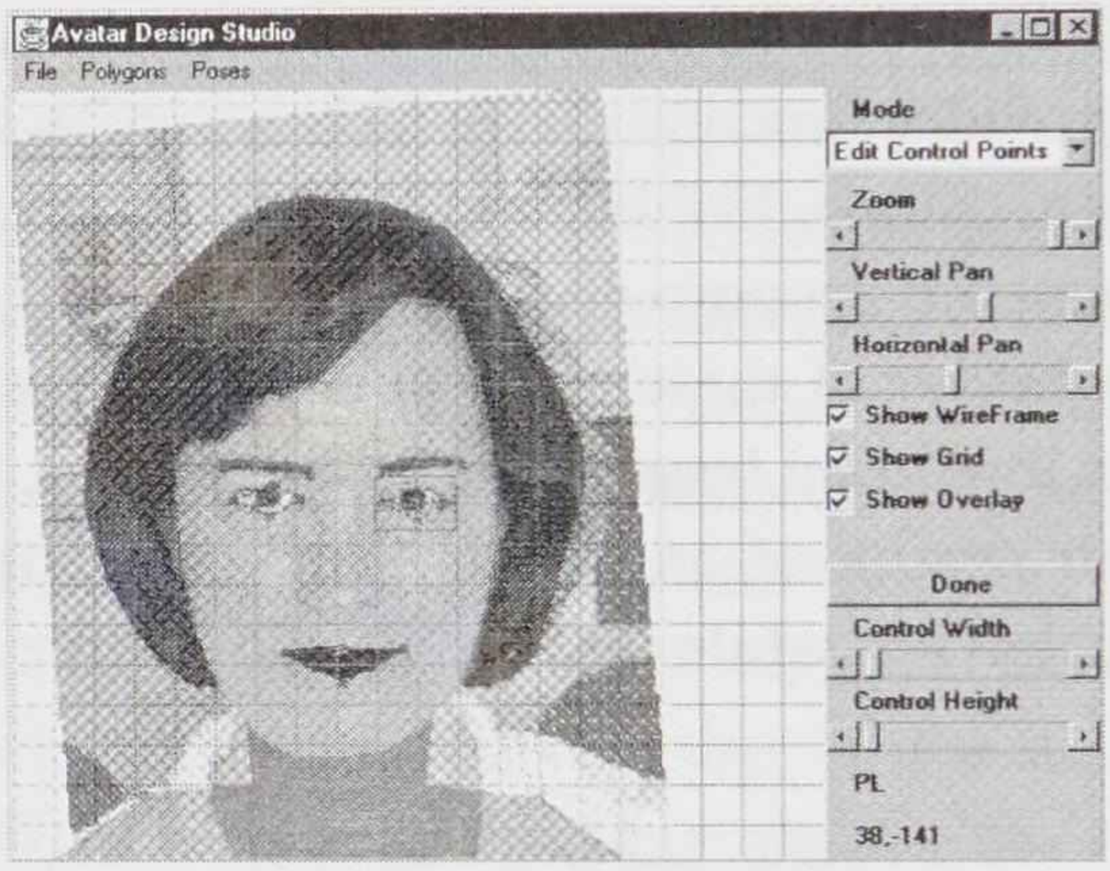


Figure 2. Avatar Authoring Studio

5.2 *Behavior*

The behavior of an avatar is specified in a scripting language. Each script defines a named behavior which consists of a set of sub-behaviors and/or primitive commands, each of which has an explicit start time (relative to the start of the script) and duration. Primitives include those for performing pre-defined poses, moving, scaling, and playing pre-recording audio files. Thus, an avatar author can create an avatar which "talks" (using their recorded voice) and uses facial expressions, gestures, and body posture (using pose primitives) all simultaneously to produce a narrative about an object in a document by creating a script with the appropriate commands.

One novel aspect of the scripting language is a set of primitives which allow an avatar to simulate user mouse-clicks on document objects, thus enabling it to give a guided tour through a series of hypertext-linked pages (e.g., for a presentation). These primitives all reference a named object on the

document (e.g., in HTML the name is that defined in the standard NAME property of the 'A' tag). These primitives are:

- MOVE_TOWARDS "object" — causes the avatar to be moved into the vicinity of the object on the page, such that it can reach the object with a hand through performance of a pre-defined pose.
- POINT_TOWARDS "object" — causes the avatar to perform one of its pre-defined poses, so that at completion its hand is on or pointing towards the specified object.
- CLICK "object" — performs the same actions as if the user had clicked on the specified object with the mouse. Accompanied by an audible "click" sound.

All of the scripts for a given avatar (pertaining to a particular document) are bundled into a compact text file, together with a reference to the file containing the definitions of the avatar's appearance. These two files are loaded from the document server for every avatar associated with a document, at the time the document is loaded by the rendering engine.

5.3 *Avatar-document binding*

The Document Avatars system works in conjunction with a hypertext document rendering system, such as an HTML web browser, which supports annotating objects on the page (e.g., word, sentence, paragraph, etc.) with hypertext links, symbolic names, and other properties. Avatars are attached to such a document by specifying a special "Avatar Reference" property associated with a page object (for example, in HTML this is accomplished by adding an "AREF" property to the 'A' tag), whose value is a list of pairs, each of which specifies an avatar description file and a behavior specified within that file. This mechanism is normally used to associate comments or narratives about the annotated object by the person represented by the specified avatar.

5.4 *User interaction*

When a document is viewed, all avatars associated with the current page are loaded and displayed "docked" in the margin of the document, as shown in Figure 3. Avatar links (as defined above) are highlighted in a color different from that used for regular hypertext links.

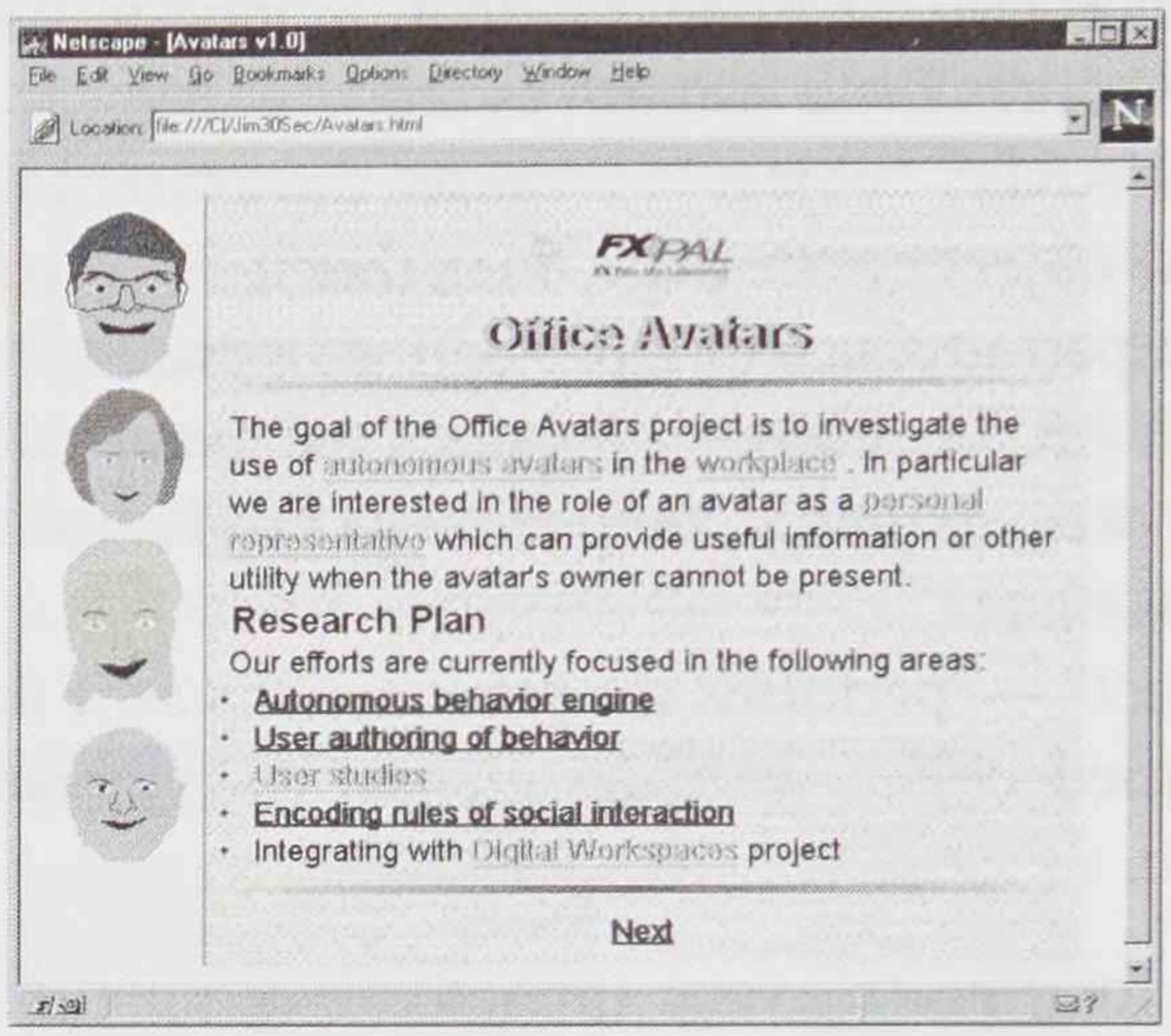


Figure 3. Document with Docked Avatars

Only the heads of avatars are shown while docked to allow multiple avatars to be displayed. Normally, when an avatar is performing a behavior it will appear on top of the document with its entire head and body rendered, as shown in Figure 4.

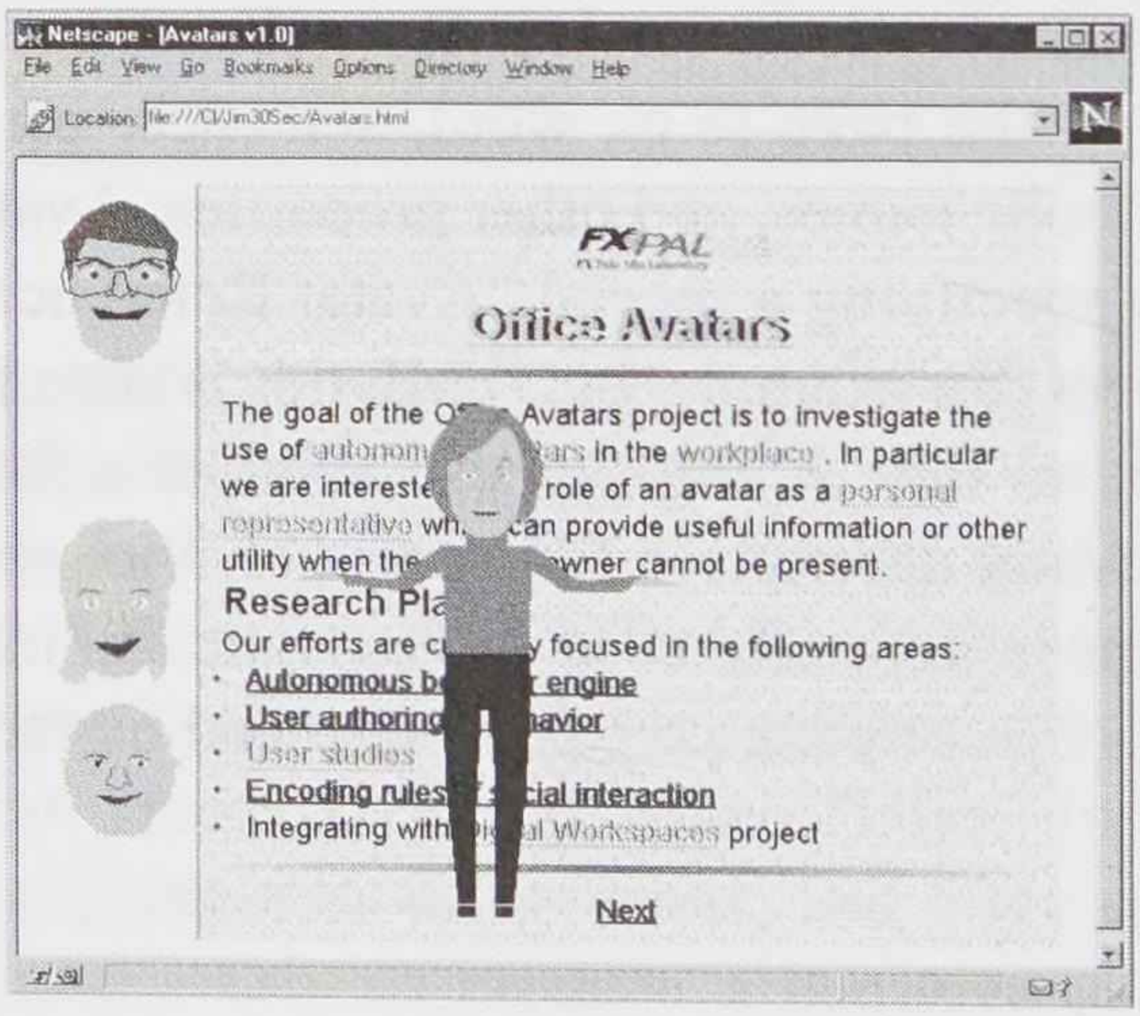


Figure 4. Avatar Performing a Behavior

The user can interact with the avatars in various ways:

- If the user clicks on an avatar link, the behavior associated with the link is performed for the specified avatar. If necessary, the avatar is moved from the margin to the general location of the annotated object, and the avatar's body is rendered, prior to the performance's beginning.
- If the user clicks on a docked avatar it will perform an "INTRO" behavior (typically an introduction or greeting).
- If the user drags the avatar from the margin onto the document it will perform a "SELECTED" behavior, if defined (typically an expression of interest).
- If the user drags the avatar over a document object for which the avatar has a behavior defined, it will perform a "DRAG_HANDLE" behavior, if defined (typically a "happy" expression).
- If the user drags the avatar over a part of the document for which the avatar does not have any behaviors defined, it will perform a "DRAG_NOHANDLE" behavior, if defined (typically a "sad" or "neutral" expression).
- If the user drags the avatar over a document object for which the avatar has a behavior defined, and releases the mouse button, the avatar will perform the associated behavior.
- If the avatar is left on the document following completion of a behavior performance, the document rendering engine will periodically cause the avatar to perform an "IDLE" behavior, if defined.

5.5 *Summary*

The Document Avatars application has been implemented as a Java applet and allows HTML web pages to be annotated with characters which can be interacted with via any Java-enabled web browser. The Avatar Authoring Studio has been implemented as a stand-alone Java application.

Although the application appears to be very similar to the Microsoft Agent (Microsoft Corporation) the two projects have very different goals and uses; Microsoft is more focused on providing high-fidelity characters as a generic OS service, and we have been more focused on user representation and character authorability. We can create a 2D character for an individual in 20 minutes, and the specification file is typically 20KB in size, enabling new characters to be quickly downloaded. In contrast, the Microsoft Agent characters are bit-mapped, requiring each frame of any animation to be hand painted,

and the resulting specification files are over 3MB. As a result, most sites using their technology use one of Microsoft's pre-defined characters rather than defining their own.

5.6 *Lessons learned*

Initial explorations suggest there are a number of general benefits to be gained using these autonomous representatives to augment documents in this way:

- Using such representatives increases the amount of information stored within a document by offering extra information when the avatar is invoked.
- This extra information does not require the viewer to leave the current page they are viewing. The user can continue to view the current page while being offered more information visually and auditorally.
- The hypertext links in the page can be ordered into a meaningful sequence by the avatar, thus offering strategies for information navigation.
- The avatar can provide a presentation which is custom-tailored to each user's needs, thus making it far more flexible than video clips or other "canned" multi-media presentations.
- In the case in which the avatar creator is also the document author, the avatar can add to the information available by providing the context of the document's construction.

Other explorations of our document-based representatives have included more public appearances in the form of participation in several presentations. In this context, seminar speakers have co-presented seminars with their autonomous personal representatives. By placing the representative on appropriate places within a projected web page, the representative is able to deliver parts of the presentation (see Figure 5). Informal discussions after these presentations have yielded considerable positive feedback from audience members. A number of studies are planned to further investigate the potential of these autonomous representatives in such settings.

One of the major research issues in extending this system is, of course, how to move from scripted behavior to autonomous behavior on the part of the representatives. Based on our experience in developing both cognitive model autonomous systems and behavior based autonomous systems,[2] our opinion is that the most difficult problem is how to achieve autonomy while still providing the proper level of end-user authorability.

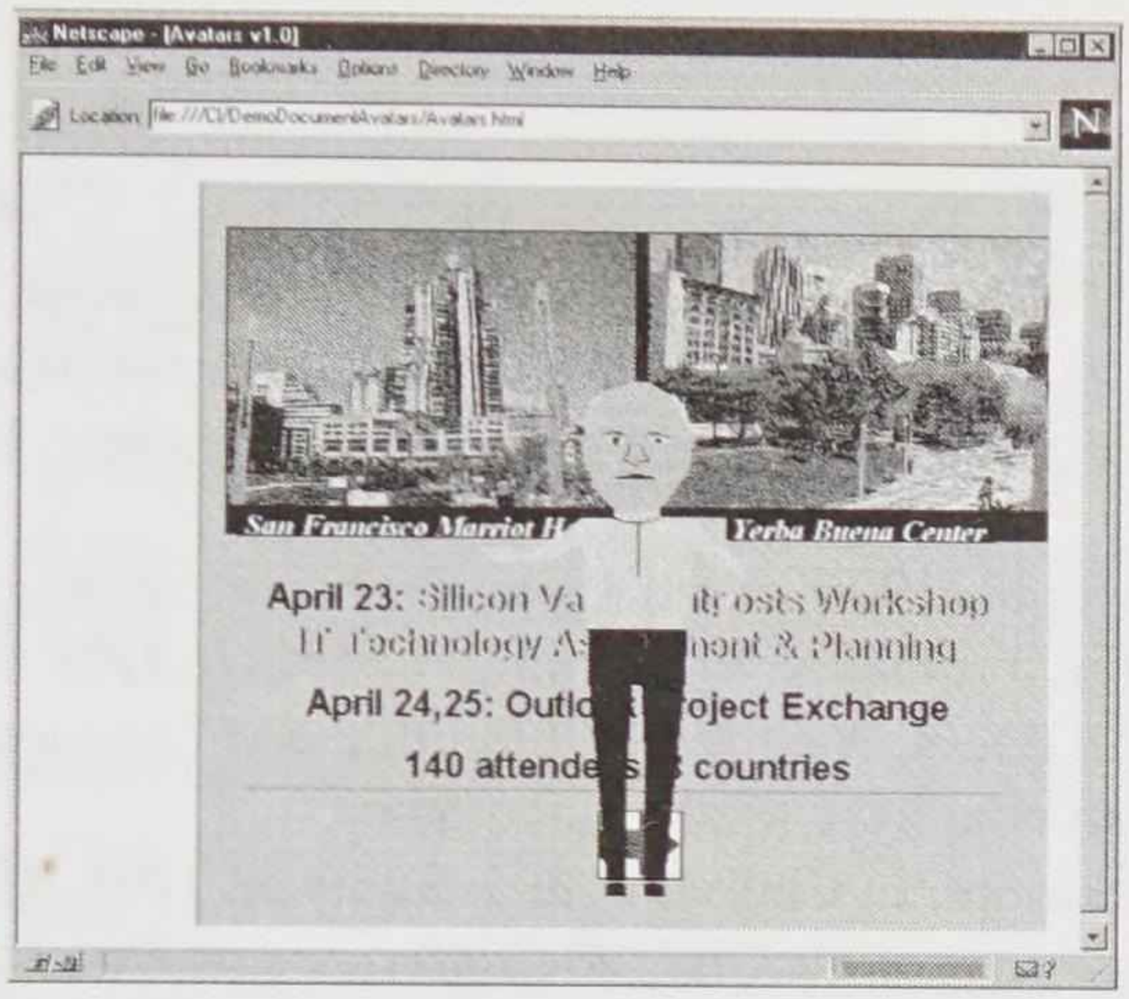


Figure 5. Example Avatar Presentation

6. Conclusions

In this paper we have described the research goals and issues in constructing autonomous personal representatives, and an initial application in which such representatives can be used as document annotations. We have several user studies planned to assess the effectiveness of such autonomous personal representatives in professional work environments. The first is to consider whether communication through these representatives aids information provision and sharing. Secondly, and at a more social psychological level, the planned studies are aimed at considering how compelling these personal representatives are and whether they have a role for building and supporting interpersonal communication and promoting camaraderie, rather than leaving individuals further isolated behind a wall of mediated interaction (Caldwell and Taha 1993; Cutler 1995).

Notes

1. For example, a quick survey of the 98 student web pages at the MIT Media Lab showed that 62% contain personal photos of the author.
2. In the Autonomous Control Logic program at Lockheed we developed a fully-operational control system for an unmanned submersible.

References

Andre, E., J. Muller & T. Rist. 1996. The PPP Persona: A Multipurpose Animated Presentation Agent. *Proceedings of Advanced Visual Interfaces*, ACM Press, 245–247.

Bickmore, T., L. Cook, & Churchill, E. 1998. Animated Autonomous Personal representatives. *Proceedings of the Autonomous Agents '98* Conference, Minneapolis, MN, May 10–13.

Blumberg, B. & T. Galyean. 1995. Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments. *Proceedings of SIGGRAPH '95*, ACM Press.

Blumberg, B. 1996. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. Unpublished doctoral dissertaion, MIT Media Laboratory.

Bly, S., L. Cook, T. Bickmore, E. Churchill, & J. Sullivan. 1998. The rise of personal web pages at work, *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*. Los Angeles: ACM, 313–314.

Caldwell, B. & L. Taha. 1993. Starving at the Banquet: Social Isolation in Electronic Communication Media. *Interpersonal Computing and Technology* (LISTSERV@GUVM.GEORGETOWN.EDU), January.

Cassell, J., T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, & Yan, H. 1998. An Architecture for Embodied Conversational Characters. *Proceedings of the First Workshop on Embodied Conversational Characters*, 1998, 21–30.

Churchill, E., S. Prevost, T. Bickmore, P. Hodgson, & L. Cook. 1998. Design Issues for Embodied Conversational Characters. *Proceedings of the First Workshop on Embodied Conversational Characters*, 1998, 149–158.

Cutler, R. 1995. Distributed Presence and Community in Cyberspace. *Interpersonal Computing and Technology* (LISTSERV@GUVM.GEORGETOWN.EDU), April, 1995.

Donath, J.S. 1999. Identity and deception in the virtual world. In Kollock, P. and M. Smith (Eds.), *Communities in Cyberspace*, Berkeley: University of California Press.

Goffman, E. 1959. *The Presentation of Self in Everyday Life*. New York: Doubleday.

Greenberg, S. 1997. Collaborative interfaces for the web. In Forsythe, C., J. Grose, & J. Ratner (Eds.), *Human Factors and Web Development,* LEA Press.

Hampson, S.E. 1988. *The Construction of Personality,* London: Routledge.

Hampson, S.E. 1992. The emergence of personality: a broader context for biological perspectives. In Gale, A. & M. W. Eysenck (Eds.), *Handbook of Individidual Differences: biological perspectives*. Chichester: Wiley.

Lathers, D. 1997. *Successful Nonverbal Communication*, Needham Heights, MA: Allen Bacon.

Maes, P. 1994. Modeling Adaptive Autonomous Agents. *Artificial Life*, Vol. 1 (Numbers 1,2), 135–162.

Microsoft Corporation. *Introduction to Microsoft Agent.* http://www.microsoft.com/workshop/prog/agent/.

Moon, Y. 1998. The effects of distance in local versus remote human-computer interaction. *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*, 103–108. Los Angeles: ACM.

Nohria, N. & R. Eccles. 1992. Face-to-face Making network organization work. In Nohria, N. & Eccles (Eds.), *Networks and Organizations*. Cambridge, MA.: HBS Press.

Noma, T. & N. Badler. 1997. Jack Presenter. Video presentation given at *Virtual2Humans Conference*, Los Angeles, June 17–19.

Olson, J., L. Covi, E. Rocco, W. Miller, & P. Allie. 1998. A room of your own: What would it take to help remote groups work as well as collocated groups? *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*, 279–280. Los Angeles: ACM.

OnLive! Technologies, http://www.onlive.com.

Prentice-Dunn, S. & R. Rogers. 1982. Effects of public and private self-awareness on de-individuation and agression. *Journal of Personality and Social Psychology*, 43: 503–513.

Rocco, E. 1998. Trust breaks down in electrical contexts but can be repaired by some initial face-to-face contact. *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*, 496–502. Los Angeles: ACM.

Schiano, D.J. & S. White. 1998. The first noble truth of cyberspace: people are people (even when they moo). *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*. 352–359. Los Angeles: ACM.

Stewart, J., E. M. Raybourn, B. Bederson, & A. Druin. 1998. When two hands are better than one: Enhancing collaboration using single display groupware. *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*. 287–292. Los Angeles: ACM.

Stone, M., K. Fishkin, & E. Bier. 1994. The Movable Filter as a User Interface Tool. *SIGCHI'94*.

The Palace, http://www.thepalace.com.

Thorisson, K. 1996a. *ToonFace: A System for Creating and Animating Interactive Cartoon Faces*. MIT Media Laboratory Learning and Common Sense Section Technical Report 96–01, April.

Thorisson, K. 1996b. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*, Unpublished doctoral dissertation, MIT Media Laboratory.

Turkle, S. 1995. *Life on the screen: Identity in the age of the Internet.* New York: Simon & Schuster.

Vere, S. & T. Bickmore. 1990. An Autonomous Agent. *Computational Intelligence*, 6: 41–60.

Vertegaal, R., V. Harro, & R. Slagter. 1998. Look who's talking: the GAZE groupware system. *Proceedings of the CHI '98 Conference on Human Factors in Computing Systems*. 293–294. Los Angeles: ACM.

Wiggins, J.A., B. B. Wiggins, & J. Vander Zanden. 1994. *Social Psychology.* New York: McGraw-Hill.

Worlds Chat. 1966. *Interactivity*, December, page 33.

Zellweger, P. 1988. Active paths through multimedia documents. *Proceedings of the International Conference On Electronic Publishing, Document Manipulation, and Typography*, 19–30

CHAPTER FIVE

"Consciousness" and Conceptual Learning in a Socially Situated Agent

Myles Bogner, Uma Ramamurthy, and Stan Franklin

1. Introduction

Wheeler notes, "orthodox cognitive science claims that situated (world-embedded) activity can be explained as the outcome of in-the-head manipulations of representations by computational information processing mechanisms" (1994). He points out the difficulty AI encountered "in moving from toy worlds to dynamic unconstrained environments," and further argues that such difficulties are inevitable within the stated paradigm. Instead, he advocates systems that "exhibit dynamical profiles comparable to those displayed by biological neural networks, and … play the same adaptive role as biological networks, i.e., to function as the control systems for complete situated agents." This view seems to us particularly relevant when applied to socially situated agents. Here we offer "Conscious" Mattie[1] as a prototype of the type of biologically motivated system Wheeler spoke of, able to interact, adapt and learn in a social environment comprised of human agents. CMattie should be equally at home in a society of agents of her own type, or in a mixed society. In this paper we will describe CMattie, a "conscious," socially situated, software agent, paying particular attention to her "consciousness" and conceptual learning mechanisms.

An *autonomous agent* (Franklin and Graesser 1997) is a system situated in, and part of, an environment, which senses that environment, and acts on it, over time, in pursuit of its own agenda. It acts in such a way as to possibly

influence what it senses at a later time. In other words, it is structurally coupled to its environment (Maturana 1975; Maturana and Varela 1980). Biological examples of autonomous agents include humans and most animals. Non-biological examples include some mobile robots, and various computational agents, including artificial life agents, software agents and computer viruses. We will be concerned with an autonomous software agent, "living" in a real world computing system.

Autonomous software agents, when equipped with cognitive (interpreted broadly) features chosen from among multiple senses, perception, concept formation, attention, problem-solving, decision making, short and long-term memory, learning, emotions, etc., are called *cognitive agents*. Though ill-defined, cognitive agents can play a synergistic role in the study of human cognition, including consciousness (Franklin 1997). In this chapter, cognitive features such as attention are used both in the folk-psychological and technical senses.

Here, we are particularly concerned with cognitive software agents that implement global workspace theory, a psychological theory of consciousness (Baars 1988, 1997). Global workspace theory postulates that human cognition is implemented by a multitude of relatively small, special purpose processes, almost always unconscious. It is a multiagent system with a society of its own. Coalitions of such processes, when aroused by novel and/or problematic situations, find their way into a global workspace (and into consciousness). This limited capacity workspace serves to broadcast the message of the coalition to all the unconscious processes, in order to recruit other processes to join in handling the current novel situation, or in solving the current problem. All this takes place under the auspices of contexts: goal contexts, perceptual contexts, conceptual contexts, and cultural contexts. Each context is itself a coalition of processes. There is much more to the theory, including attention, learning, action selection, and problem solving.

We will refer to cognitive agents that implement global workspace theory as "*conscious" software agents*. "Conscious" software agents are domain-specific entities; very little of their architectures is domain-independent. They adapt and learn by reacting to the changes in their domain, and through their interaction with other agents in their domains, be they human or artificial. Due to this extensive interaction, "conscious" software agents tend to be social creatures, and exhibit some socially situated intelligence.

CMattie is such a "conscious" agent (Franklin and Graesser In Press).

Designed for a specific, narrow domain, she functions in an academic setting, "living" in a UNIX-based system. She gathers information from humans regarding seminars and seminar-like events such as colloquia, theses defense, etc. Using this information, she composes an announcement of the next week's seminars, and mails this announcement weekly to members of a mailing list that she maintains, again by email interactions with humans. CMattie uses short-term, intermediate and long-term memories. Her emotion module enables her to react to both internal and perceived events. Her sense of self-preservation underlies her concern about her resource needs and about the status of the UNIX-based system in which she "lives." By interacting with seminar organizers, CMattie learns new concepts and behaviors. Due to the nature of her domain, this learning occurs mainly through case-based reasoning. Such learning mechanisms contribute to CMattie's socially situated intelligence (SSI). She learns to react differently in different social situations.

Following the tenets of the action selection paradigm (Franklin 1995) as expanded into design criteria (Franklin 1997), CMattie is designed using a multiplicity of artificial intelligence mechanisms. Her modular architecture, as illustrated in Figure 1, implements and integrates these several diverse mechanisms. They include behavior networks (Maes 1990) for action selection, sparse distributed memory (Kanerva 1988) for long-term, associative memory, pandemonium theory (Jackson 1987) for agent grouping, Copycat architecture (Mitchell 1993; Hofstadter and Mitchell 1994) and natural language understanding (Allen 1995) for email comprehension, and case-based memory (Kolodner 1993) for intermediate term, episodic memory. Each of these mechanisms has been significantly extended in order to merge with the others, and to meet the needs of this domain. CMattie is the first software agent intended as an implementation of global workspace theory.

CMattie is different from other email and scheduling agents. For example, the Calendar Agent automates a user's scheduling process by observing the person's actions and receiving direct feedback (Kozierok 1993). The Maxims system is an email filtering agent which learns to process a user's incoming mail messages (Lashkari 1994). These two systems employ other agents that collaborate to overcome the problem of learning from scratch. Re:Agent is an email management system (Boone 1998). This agent routes email to handlers that delete, download, sort, and store these messages on palmtop computers and pagers. Re:Agent learns the emails' features in order to learn how to appropriately classify the messages. The Visitor-Hoster sys-

tem is aimed at helping a human secretary organize a visit to an academic department (Sycara 1994). The secretary is presented with a user interface where she inputs relevant information to the agent about the incoming visitor. The agent then plans the visit, and returns to the secretary for confirmation. In addition to differences in tasks, CMattie's architecture, method of communication, and degree of autonomy make her relatively unique among these types of agents. Her architecture combines numerous artificial intelligence techniques to model the human mind. She communicates entirely via the natural language found in email messages. CMattie is designed to fully function as a seminar coordinator.

CMattie has several drives, some corresponding to her tasks (sending seminar announcements, reminding organizers to send information, and acknowledging messages). These drives are explicitly built into the agent, and operate in parallel. Some drives vary in urgency, an extension to Maes' work. For example, the urgency level for sending out a seminar announcement will increase as the time to send the announcement approaches. Each drive activates behaviors that work to fulfill the drive.

Behaviors in CMattie (Song 1998) correspond to global workspace theory's goal contexts. Each behavior has an activation level affected by drives, other behaviors adjacent to it in the behavior net, internal conditions, and environmental inputs (the perception registers' contents). Only one behavior can be active at a time. A behavior's activation is spread to those behaviors that can fulfill its unmet preconditions and to behaviors whose preconditions can be satisfied by this behavior. Each behavior can thus be considered part of a behavior stream. For example, there's a behavior stream that composes the seminar announcement. One behavior in that stream might fill the Cognitive Science Seminar's portion of the seminar announcement.

CMattie's emotions play two roles (McCauley and Franklin 1998). First, emotions indirectly affect a behavior stream's activation level by affecting the strength of drives. Emotions allow CMattie to be pleased about sending out a seminar announcement on time and to be anxious about an impending system shutdown. In these cases, emotion might increase a behavior stream's activation level since it is pleasing for CMattie to complete these streams promptly. Second, emotions influence the suggested actions that are the output of sparse distributed memory. Therefore, CMattie may be more or less inclined to actively pursue a suggested action based on the action's associated emotional level.

2. Overview of an Architecture for Supporting SSI

CMattie's architecture is quite complex as seen in Figure 1. Our references point to several papers where specific modules of the architecture have been described in depth (Ramamurthy, Bogner, and Franklin 1998; Bogner 1998; McCauley and Franklin 1998; Zhang, Franklin, and Dasgupta 1998; Ramamurthy, Franklin, and Negatu 1998). Here we present a brief overview of CMattie's architecture so that the reader can follow our discussion.

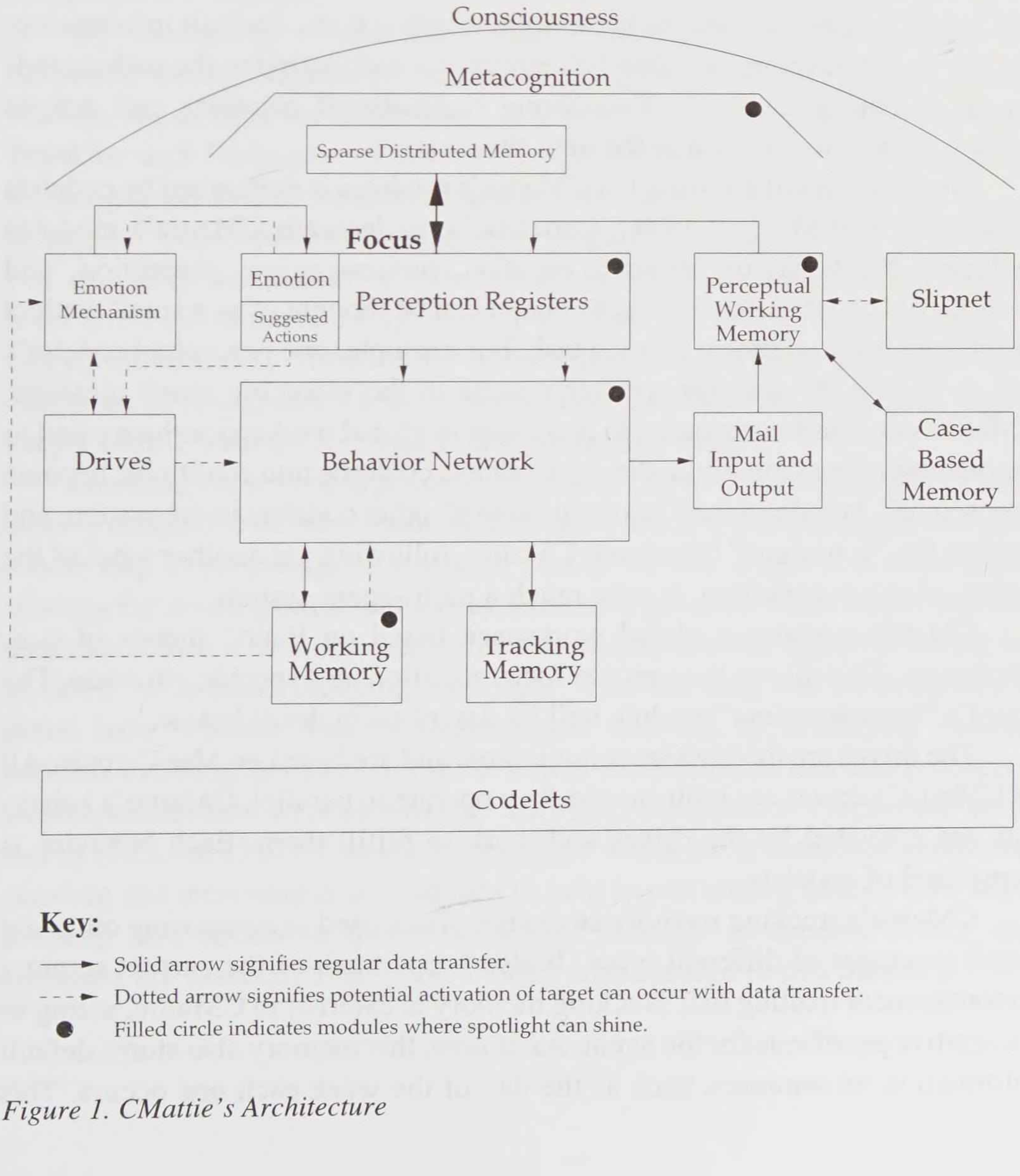


Figure 1. CMattie's Architecture

Incoming email messages are first received by the mail input portion and are then moved to the perception module. CMattie's sensory data are, for the most part, the incoming email messages she receives. Perception for the agent occurs when she comprehends such a message. Comprehended messages are placed in the focus. CMattie's perception is described in more detail in the next section.

The focus serves as an interaction point for several of CMattie's modules, including sparse distributed memory. Sparse distributed memory is a content addressable memory that serves as long-term, associative memory for CMattie (Anwar and Franklin, forthcoming). This memory stores the contents of the perceptual registers as well as her emotions and actions. Default information, such as time and room can often be recovered, contributing to the understanding of incoming messages. Recovering remembered emotions and actions helps with action selection in the new situation.

The real work of almost all of CMattie's modules is performed by codelets (Hofstadter and Mitchell 1994). Codelets lie underneath CMattie's modules including her behavior network, emotion, metacognition, perception, and portions of "consciousness". Each codelet can be thought of as a small distinct agent designed to perform a single task. For example, one perceptual codelet's task is to find the seminar speaker's name in the incoming email message. CMattie's codelets correspond to processes in global workspace theory and to the demons of pandemonium theory. Codelets coalesce into coalitions, become "conscious", broadcast their information to all other codelets in the system, and receive the "conscious" broadcast. CMattie, following yet another tenet of the action selection paradigm, is very much a multi-agent system.

CMattie contains a global workspace based on Baars' theory of consciousness. This allows the agent to focus attention on a specific situation. The agent's "consciousness" module will be described in detail below.

The drives are the high level motivators and are based on Maes' goals. All of CMattie's drives are built-in, and they operate in parallel. CMattie's behaviors are activated by the drives and work to fulfill them. Each behavior is comprised of codelets.

CMattie's tracking memory stores templates used in composing outgoing email messages of different types. It also keeps track of the current seminar announcement mailing list. Tracking memory is external to CMattie, acting as a cognitive prosthesis for the agent. As of now, this memory also stores default information on seminars, such as the day of the week each one occurs. This

function will probably be subsumed by associative memory.

All outgoing messages are composed in the composition workspace. Message composition consists of filling the fields of an outgoing message template. The information used to fill these fields comes from the perception registers and any of associative, case-based, or tracking memories. A current seminar announcement template is always being generated in the composition workspace. As new information is perceived and placed in perception registers, the template fields are filled. When a seminar announcement is moved to mail output and mailed, a new announcement template is placed in the composition workspace.

By monitoring what is in "consciousness", the activation of drives, emotional states, parameters in the behavior network, and the perception module, metacognition keeps track of CMattie's internal conditions (Zhang, Franklin and Dasgupta 1998). Using a classifier system (Holland 1986), metacognition makes inferences about CMattie's state. If necessary, it can influence "consciousness", perception, learning, and the behavior network. For example, metacognition can change the behavior network's activation level threshold to make the agent more goal-oriented or more opportunistic. It can cause voluntary attention by influencing the activation levels of certain coalitions of processors. Metacognition plays the role of an overseer, trying to keep CMattie's action selections on a productive track.

Learning via several types of mechanisms allows CMattie to become more closely coupled to her environment. She can learn new behaviors, for example, a new step in preparing for a system shutdown. She might also learn a new strategy for sending out reminders to seminar organizers. Much of her learning uses case-based reasoning. She learns new concepts in her slipnet allowing her to better understand incoming messages. This learning will be described in detail later. CMattie creates (learns) new codelets by modifying existing codelets enabling her to perform the newly learned behaviors and perceptual techniques. Coalitions of codelets are learned via association a la pandemonium theory (Jackson 1987). This allows the agent's codelets greater ease in communicating and recruiting other codelets to help in performing tasks. Associative learning also occurs in sparse distributed memory as actions, events, and emotions are associated with one another when placed in this memory.

3. Perception

The perception module in CMattie (Figure 2) was inspired by and can be thought of as an extension of the Copycat architecture (Mitchell 1993). Copycat is based on the premise that analogy-making is a process of high-level perception, and that analogy-making lies at the core of understanding. Copycat makes and interprets analogies between situations in a microworld of letter-string analogy problems. Copycat's domain is predefined and fixed; therefore, there is no learning. Since CMattie "lives" in a dynamic domain, her perceptual learning mechanism enables her to perceive this dynamism.

CMattie's perception involves building instances of known concepts in her domain; learning, detecting and creating new concepts; and making appropriate relations between those concepts. Her perception module consists of mail input and output, the slipnet, working memory and case-based memory.

Mail Input and Output. These provide CMattie's interface to her domain. Using this unit, she receives and sends out email messages related to seminars, seminar-like events such as colloquia, and maintenance of the recipient mailing list. Mail input and output can process more than one email message at a time, enabling the perception module to perceive and understand emergency events in CMattie's world. This aids in maintaining her sense of self-preservation as she proactively reacts to her changing resource needs. She immediately reacts to the status of the UNIX-host system wherein she "lives".

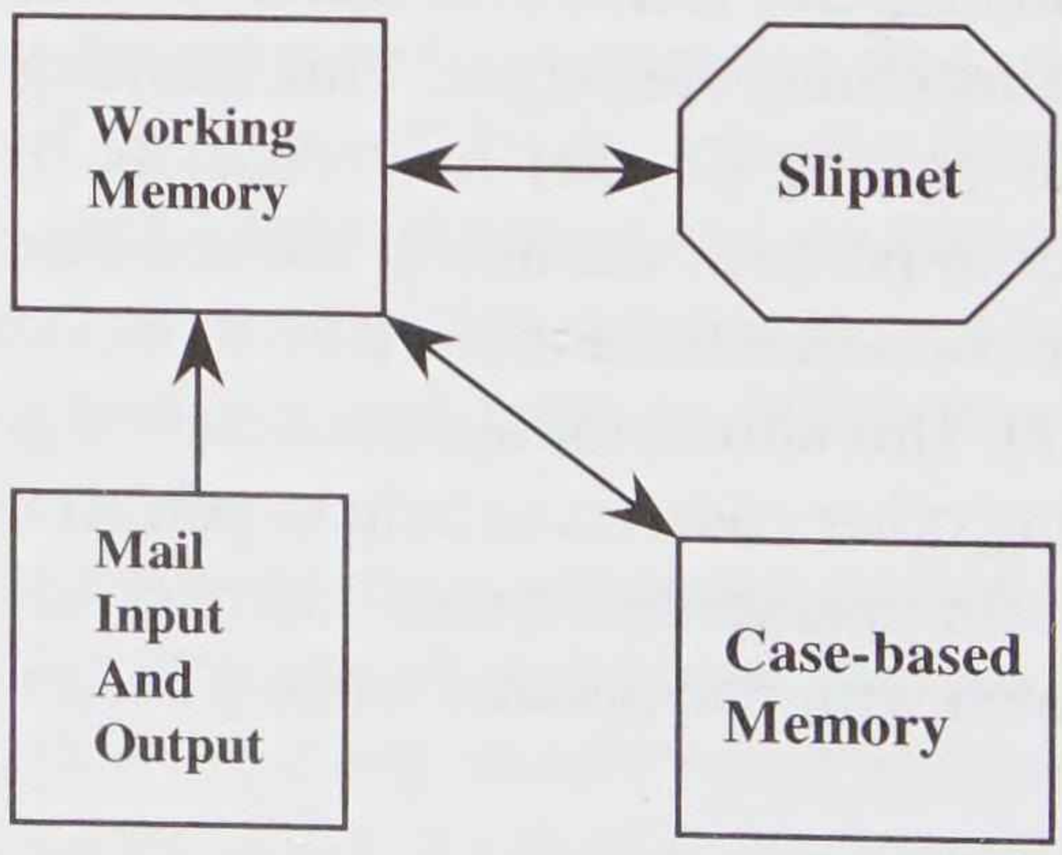


Figure 2. CMattie's Perception Module

Slipnet. The slipnet is a network of nodes and links representing CMattie's permanent perceptual concepts. A concept has a core and a set of features representing its basic characteristics. In a given context, a feature might have a specific value. In CMattie, concepts are often defined by a region of nodes and links in the slipnet. Each of the concepts in the agent may be an individual node or a group of nodes. The various nodes are connected to each other through weighted links.

One of the built-in concepts in the perception module is the Seminar concept with the following features:

Name of the seminar
Organizer of the seminar
Location where the seminar is to be held
Date of the seminar
Day of the week of the seminar
Time at which the seminar is to be held
Speaker of the seminar
Title of Talk for the seminar
Periodicity of the seminar

Name and *Day* are features of the Seminar concept, and they are concepts themselves, each with a separate set of features. For example, as shown in Figure 3, the *Day* node in CMattie's slipnet is a concept with nodes *Monday*, *Tuesday* as its features. The seminar concept is deeper than the *Name* and *Day* concepts and, therefore, has a higher *depth value* than those two concepts. Depth values aid in the assignment of node activation level.

Each node in the slipnet has one or more codelets associated with it. When an email message is received by mail input, these codelets aid in understanding the message, which is written in natural language (Zhang et al. 1998). They recognize relevant words and phrases in the received message, and send activation to the appropriate slipnet nodes. A corpus of email messages collected for two years contributed to the building of the slipnet.

Working Memory. This memory holds the contents of the incoming email message. It also holds the perception process' intermediate results, as codelets associated with slipnet nodes operate inside working memory to understand the received email message. The most significant inference made in this process is the categorization of the *type* of the incoming message.

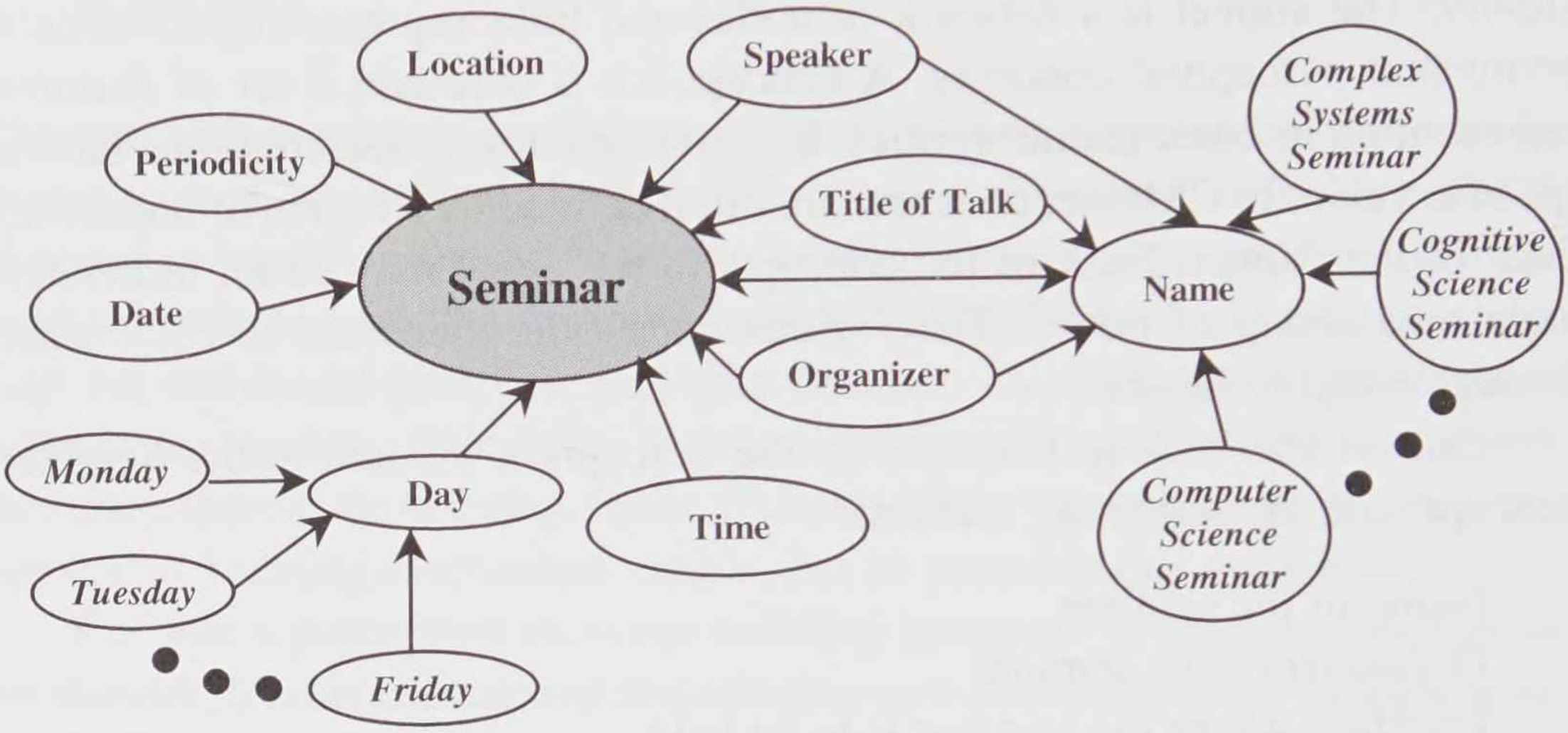


Figure 3. Segment of the Slipnet in the Perception Module

Case-based Memory. Case-based memory constitutes CMattie's episodic memory. In it she stores the sequences of email messages that form episodes. This allows her to relate new events to similar past events. She understands these past events using her built-in domain knowledge. Case-based memory aids her in learning new slipnet concepts through case-based reasoning. This memory acts as an intermediate term memory, and the information stored there is used to learn domain knowledge.

3.1. *Perception process*

When an incoming message is understood, every significant word or phrase has been given a field name, and the *type* of the email message has been inferred. This information is then transferred by perceptual codelets to the perception registers in the focus. Some of the perception registers are Name, Organizer, Location, Date, Day, Time, Speaker, Title-of-Talk, Periodicity, and Message Type. Other perception registers hold previously unencountered words and phrases that occur in the received email messages and that might be relevant. The perception process is complete when the *type* of the received message has been inferred and the understood information regarding the received message has been transferred to the perception registers.

4. Bringing the Focus to "Consciousness"

Many of CMattie's components use information from the focus (Figure 4). This section describes how the focus is used to bring perceived information into "consciousness". The focus includes four vectors: the perception registers, the output of case-based memory, the output of sparse distributed memory, and the input to both case-based memory and sparse distributed memory. The perception module places the components of the understood email message into the perception registers. That constitutes the current percept. Next, sparse distributed memory is read with the current percept as the address. Also, case-based memory is read with the same address. These reads are designed to gather the information most relevant to what was just perceived.

A "consciousness" codelet is one whose function is to bring specific information to "consciousness" (Bogner 1998). In particular, after the memory reads, perceptual "consciousness" codelets bring information from the focus to "consciousness". One such "consciousness" codelet is associated with each of the perception registers and carries the specific piece of perceived information from that register. For example, one codelet carries the speaker's name, and another carries the seminar's time.

Specific "consciousness" codelets spring into action when the information in the perception registers is relevant to them. For example, if what is perceived is a request to be removed from the seminar announcement mailing

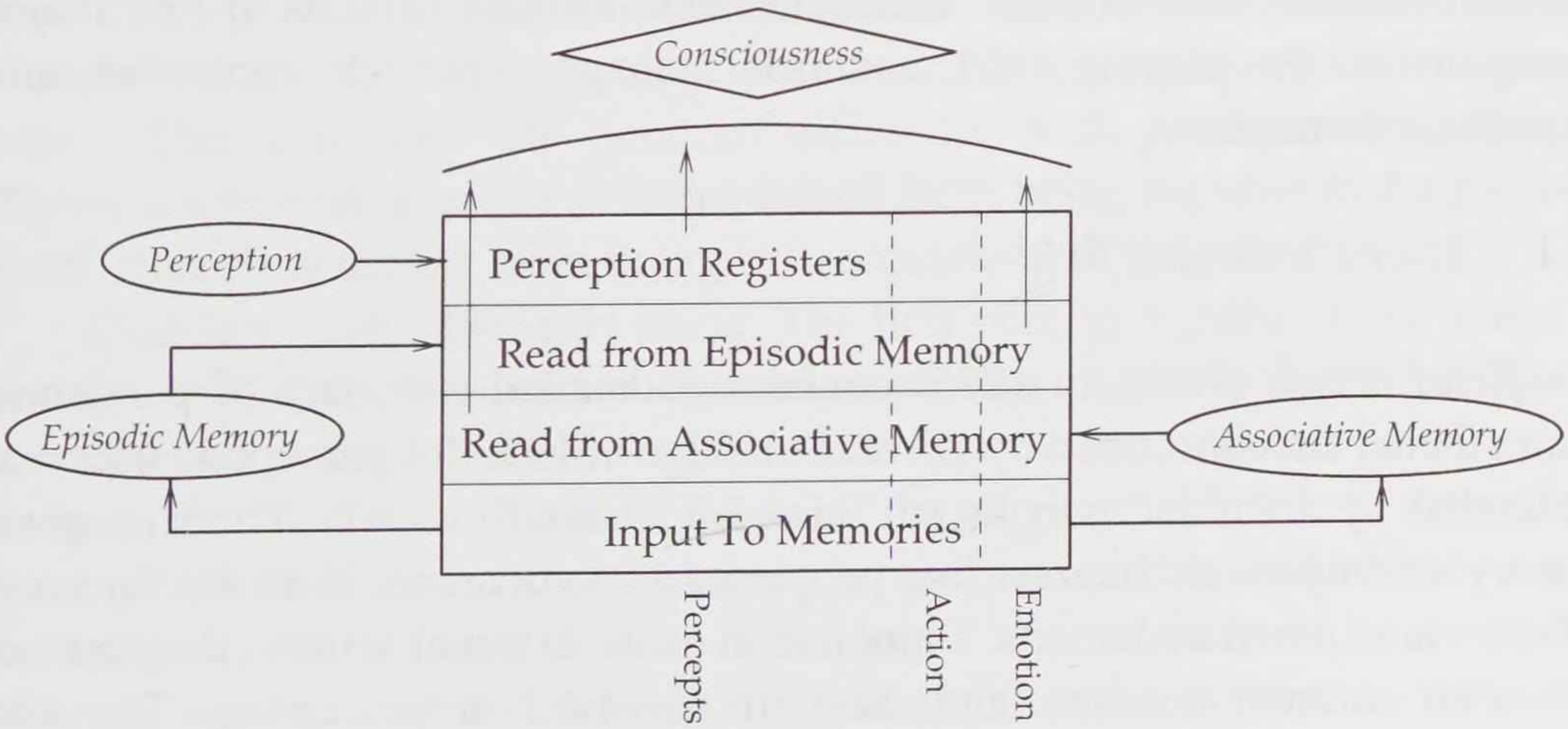


Figure 4. CMattie's Focus

list, the "consciousness" codelet which carries a person's email address becomes active. It then joins the playing field on its quest for "consciousness".

In addition, some "consciousness" codelets check for conflicts amongst the relevant items returned from the percept and the memory reads. For example, a conflict occurs if the perceived place, time and date for the Cognitive Science Seminar are the same as case-based memory's output of these same features for the Graph Theory Seminar. The "consciousness" codelet recognizing the conflict joins the playing field and raises its activation level. Since it is associated with the other perceptual "consciousness" codelets, the coalition manager (described below) groups them together to form a coalition.

5. "Consciousness"

CMattie's global workspace gives the agent several important performance features. It allows for coalitions of codelets to gain attention. Information about these codelets is broadcast to all of the agent's other codelets. Recipients of this broadcast become active themselves if enough of the information is understood, and if it is applicable. In this way, the broadcast recipients have the potential to contribute towards solving the problem raised by the "conscious" coalition. This broadcast also allows metacognition a view of the events taking place in the system. Learning also uses the information in "consciousness" to learn to associate codelets as a coalition. In addition to "consciousness" codelets, the "consciousness" module consists of four major components: the playing field, coalition manager, spotlight controller, and broadcast manager.

5.1 *CMattie's playing field*

Artificial Minds (Franklin 1995) contains a detailed summary of pandemonium theory first described by Oliver Selfridge in 1959 for perceptual uses and extended by John Jackson to an "idea for a mind" (1987). Pandemonium theory's components interact like people in a sports arena. Both the fans and players are known as demons. Demons can cause external actions, they can act on other internal demons, and they are involved in perception. The vast majority of demons are the audience in the stands. There are a small number of demons on the playing field. These demons are attempting to excite the fans.

Audience members respond in varying degrees to these attempts to excite them, with the more excited fans yelling louder. The loudest fan goes down to the playing field and joins the players, perhaps causing one of the players to return to the stands. The louder fans are those who are most closely linked to the players. There are initial links in the system. Links are created and strengthened by the amount of time demons spend together on the playing field and by the system's overall motivational level at the time.

CMattie uses pandemonium theory's notion of a playing field. A collection of codelets which act as demons are instantiated when the program first runs. Each of these is a generator codelet of a specific codelet type. If a codelet of one of these types is to become active as a result of having received information broadcast from the "conscious" coalition, the appropriate generator codelet instantiates a copy of itself with the relevant information. This allows for multiple codelets of the same codelet type to run in parallel, each working with different information. These generator codelets can be considered fans in pandemonium theory's arena.

All codelets, other than these generator codelets and the "consciousness" codelets which have not detected relevant information, are considered active and are performing their functions. These active codelets are pandemonium theory's players on the playing field. The playing field is a shared space in memory; all active codelets exist in this shared memory space.

Codelets on the playing field may be associated with one another. Some of these links are built-in. For example, codelets underlying the same higher level concept, such as a behavior, are likely to be associated with one another. Codelet associations also develop when codelets are together in "consciousness". This illustrates one point of difference with pandemonium theory. There, association arises or is strengthened from being together in the playing field. Here, it is mostly from being in "consciousness" together.

Codelets have a two-part name. The first portion signifies from where a codelet on the playing field is derived, such as a particular behavior. Since there can be multiple codelets of the same type active, codelets also carry a unique identification number. Codelets on the playing field have an activation level, which may come from the higher level construct from which they were instantiated, for example from a behavior, a slipnet node or an emotion. "Consciousness" codelets provide their own activation. Activation normally decays over time. The activation level of codelets is an important factor in deciding which coalition gains "conscious" attention.

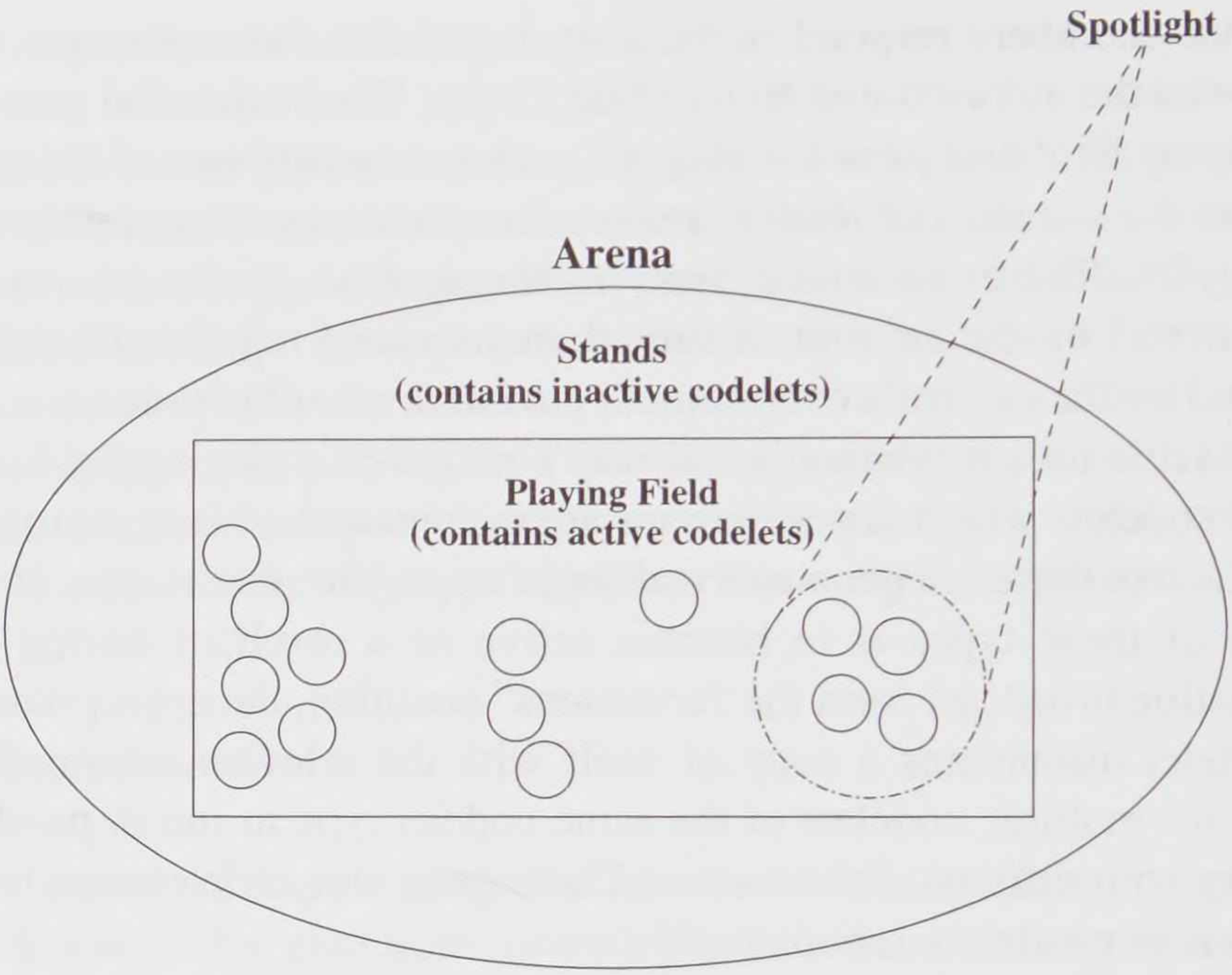


Figure 5. CMattie's Playing Field

Figure 5 illustrates CMattie's playing field. Two components of her global workspace implementation, the coalition manager and the spotlight controller, play important roles on the playing field.

5.2 *Coalition manager*

The coalition manager groups active codelets into coalitions, and keeps track of them. To make coalitions, the coalition manager groups codelets according to the strength of the associations between them. If a collection of codelets is associated above a certain threshold level, these codelets are considered to be in a coalition. Coalitions are capped on average to a maximum of seven codelets. Therefore, all of the codelets associated with a single higher level concept may or may not be in the same coalition.

The playing field provides an active dynamic environment. The activation levels of codelets continually decay. Newly activated codelets join existing coalitions. Codelets leave one coalition and possibly join another. Codelets leave the playing field when their actions are complete. Due to this dynamic environment, the coalition manager must continually and efficiently survey the playing field to keep its record of coalitions up to date.

5.3 *Spotlight controller*

The spotlight controller determines which coalition becomes "conscious". It calculates the average activation level of each of the coalitions by averaging the activation levels of the coalition's codelets. The spotlight shines on the coalition with the highest average activation level. Average activation among a coalition's codelets, not the total activation, is taken to prevent larger coalitions from having an advantage over smaller ones. In the same way as the coalition manager's, the spotlight controller's domain is extremely dynamic. Here are some instances. An activation level goes to zero when an instantiated codelet's work is complete. A "consciousness" codelet may greatly increase its activation when it, say, finds a conflict. A behavior being chosen sends new activation to each of its underlying codelets. And so on.

5.4 *Broadcast manager*

Once the spotlight controller has determined a "conscious" coalition, it notifies the broadcast manager who is responsible for gathering information from the "conscious" coalition, and sending it to all of CMattie's codelets. As in global workspace theory, messages are small and understood by only some of the agent's codelets.

Specifically, from the "conscious" coalition the broadcast manager gathers objects labeled for broadcast. These objects contain information needed for specifying the current novelty or problem. This information is then broadcast to all of CMattie's generator codelets.

In addition to being broadcast, information gathered from a coalition is placed on the blackboard, implemented as a shared memory space. This blackboard, containing at most the last seven broadcasts, or the last seven items in "consciousness", serves to implement short-term working memory. Codelets can poll this blackboard and search for parameters which they understand.

6. "Consciousness" as a Facilitator for Learning

The "conscious" broadcast recruits codelets that understand the message and for which it is relevant. This causes their activation to increase, motivating them to begin performing their respective tasks. These tasks might include

activating their overlying higher construct, say a behavior, an emotion, a slipnet node, or a learning mechanism. Figure 6 illustrates the significant role of "consciousness" in perceptual learning, highlighting global workspace theory's premise that "consciousness" is sufficient for learning. This section focuses on the perceptual learning that results from the "conscious" broadcasts.

CMattie has a limited number of seminars already defined in her slipnet. She "knows" about these seminars through the built-in seminar concept and its features. In particular, she knows that:

- A seminar is held once a week
- It has an organizer and a name
- Each week, there might be a different speaker
- It has a different title-of-talk
- It is usually held at the same location, on the same day of the week, and at the same time.

Suppose a seminar organizer sends her a message announcing a seminar with a seminar name that she has never seen before. CMattie attempts to treat such a message in a way similar to seminars that she already knows. The learning mechanism described here is based on the premise that any agent, including

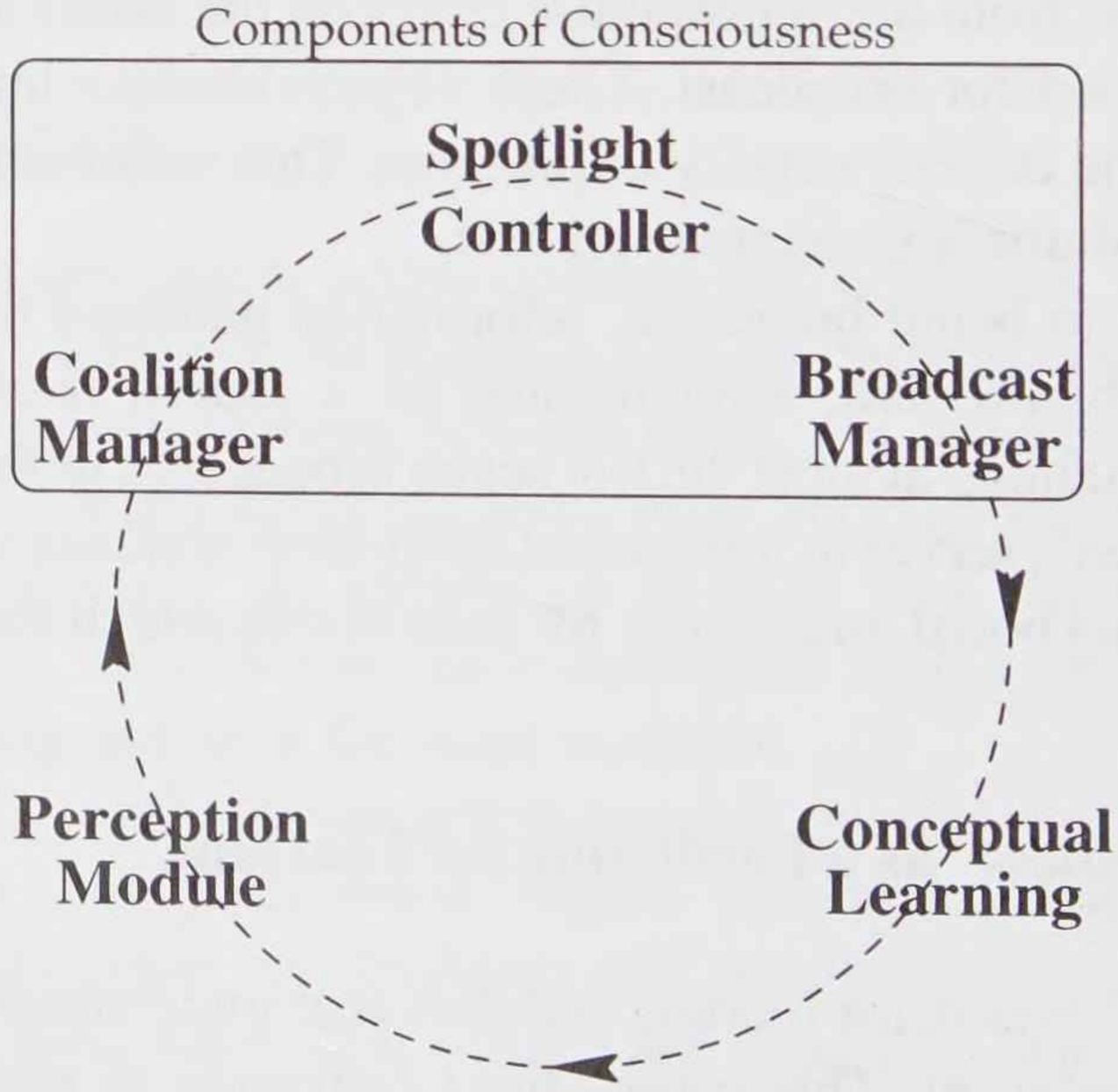


Figure 6. "Consciousness" as a Facilitator for Conceptual Learning

humans, learns based on what it already knows. When the message understanding mechanism attempts to understand this message, the agent recognizes that it is an *initiate-seminar-message* for a seminar, but the *name* of this seminar is not part of the built-in knowledge. This information is placed into the perception registers, brought to "consciousness", and broadcast. CMattie has codelets that understand this broadcast and can activate behavior streams that act to converse with the sender of the message to determine if the sender wishes to initialize a new seminar. She sends an acknowledgement to the sender stating that a new seminar with that seminar name will be initialized, with the sender as its organizer and requesting confirmation. Reinforcement of a sort is provided to CMattie by the response she might or might not get. Based on the "conscious" broadcast of this feedback, if any, a new slipnet node is created for this seminar name, and it is linked to the *name* node, which is also a feature of the seminar concept. When this new node is generated, the associated codelets for it are generated as well, a quite straightforward process. The new codelets are based on similar, existing codelets for the other *name* nodes. Once the process is complete, CMattie has understood the incoming confirmation message, and the perception module sends the relevant fields to the perception registers.

The second type of learning that takes place in the perception module occurs when CMattie learns concepts that are not completely identical to the built-in seminar concept, but slightly different from it. In her domain, colloquia, dissertation defenses, dissertation committee meetings, and faculty meetings, all fall into this category. This second learning mechanism is based on viewing every new situation in terms of a previously solved problem (analogy-making). When CMattie receives a message about such a non-seminar event, say a dissertation defense, she treats it as a *speaker-topic message* for a seminar. This understanding is disseminated through "consciousness". The agent sends an acknowledgement to the sender stating that she is initializing a new seminar by the name "Dissertation Defense Seminar" with the sender as organizer. This misunderstanding can be expected to result in one or more of the following events, depending on the sender.

– The acknowledgement elicits a negative response from the sender, starting an episode. The resulting "conversation" between CMattie and the sender is stored in case-based memory. This episode provides information that allows CMattie, even with her limited natural language understanding, to learn that dissertation defense is similar to the seminar

concept, but with slightly different features. In this case, the *periodicity* feature has a different value. CMattie learns this through case-based reasoning.

- The sender ignores the acknowledgement, and CMattie includes the Dissertation Defense Seminar in her weekly seminar announcement. In this case nothing is learned at this time, but perhaps later.
- CMattie includes the Dissertation Defense Seminar in seminar announcement. This action is likely to elicit a negative response from the sender, starting an interaction with CMattie. This episode again is stored in case-based memory to aid her in learning what a dissertation defense is.
- The sender might also ignore the incorrect weekly announcement, but respond to the reminder sent by CMattie the following week, when she doesn't receive a *speaker-topic message* for the Dissertation Defense Seminar. This, again, generates an episode allowing CMattie to learn about a dissertation defense.
- The sender ignores all the reminders. In this case, ignoring the reminders itself acts as feedback to CMattie, giving rise to a new concept that is similar to the seminar concept with its periodicity feature modified.

Regardless of which of these scenarios occur, CMattie eventually learns a new concept called *dissertation defense* that is closely related to the seminar concept. Note that this conceptual learning takes place through the internal interaction between "consciousness" and perception. Each of these possible situations becomes "conscious" before any changes to the perception module occur. Also note the crucial role in this conceptual learning played by external social interaction with a human.

How is all this done? There are two main capabilities. First, CMattie has codelets in her slipnet that look for words and phrases that she has not previously encountered. The perception module tracks such new words and phrases that occur with any regularity by keeping statistics and recognizing novelty. This aids in CMattie's natural language understanding. Second, even with her limited natural language understanding, CMattie can understand messages from organizers that have negative connotations. Her slipnet has nodes and codelets that detect words and phrases with such negative connotations.

Consider a possible path CMattie might take during her conceptual learning. CMattie misunderstands the first dissertation defense message, and sends an acknowledgement to the organizer for a Dissertation Defense Semi-

nar. Suppose the organizer responds with a negative message saying, "It is not a seminar, but a dissertation defense." CMattie understands the negative connotation in "not a seminar" and the repeated occurrence of the phrase "dissertation defense" activates her questioning capability to send a message to the organizer with the question, "How does a dissertation defense differ from a seminar?" The organizer might reply with a simple explanation such as, "A dissertation defense is like a seminar, but it might not occur regularly" or "Dissertation defenses do not ordinarily occur every week". CMattie understands the negative connotations in relation to the words "regularly," "every week" or "weekly". These are keywords in the slipnet related to the *periodicity* feature of the seminar concept. CMattie uses her case-based memory and natural language understanding to reason that a dissertation defense has a periodicity different from that of a seminar. This interaction with the organizer and her reasoning effects the creation of a new concept, *dissertation defense*, in the slipnet with related codelets that search for it in future messages. This "conscious" learning enables her to correctly perceive and understand a dissertation defense message when she next encounters one. Of course, CMattie must also learn to behave differently when faced with a dissertation defense message than she does with a seminar message. That is a subject for another chapter.

A trace of this learning stored in her case-based memory serves to enhance her case-based reasoning capabilities. Later, CMattie might encounter a colloquium message, and in response to her incorrect acknowledgement of a Colloquium Seminar, be told that "It is a colloquium, not a Colloquium Seminar." CMattie's case-based reasoning depends on (a) past experiences she has had, and (b) her ability to understand new situations in terms of her past experiences. She recalls her experience with the first dissertation defense message from her case-based memory, and reasons that *colloquium* might be similar to *dissertation defense*. She sends a message to the organizer with the question, "Is a colloquium similar to a dissertation defense?" Her understanding and reasoning, based on the organizer's reply to her query, aid her in learning about colloquia.

Thus CMattie's conceptual learning is socially situated through her "conscious" interaction with seminar organizers. This interaction enables her to acquire a measure of socially situated intelligence.

7. Criteria for Performance Analyses

Design and development of CMattie has been ongoing for several years. As of this writing, CMattie is actively being implemented. CMattie is a successor to the successful Virtual Mattie. CMattie's task set is a superset of VMattie's, a "non-conscious" autonomous agent. Here we present a portion of VMattie's test results to illustrate what tasks CMattie is expected to perform (Song 1998).

VMattie was tested over a period of four weeks. These tests were designed to simulate real world settings. During testing, she received 55 messages comprised of 10 message types. The majority of messages received fell into the categories of Seminar Initiation, Speaker-Topic, Seminar Conclusion, and Add to Mailing List. She received 5 messages which were irrelevant to her domain.

VMattie was able to correctly fill all of the perception registers for 96.4% of the messages she received. She chose the date of seminar and title of talk incorrectly for only two Speaker-Topic messages as two words were collapsed together without a space in the incoming messages. Even with this misperception, she correctly composed acknowledgement messages and sent them to the senders of each received message.

The behavior network used this perceived information to generate seminar announcements. VMattie was 100% accurate in generating and sending out the seminar announcements. This included correctly recovering missing information from her tracking memory for default values with full accuracy. VMattie was able to correctly change her mailing list upon receipt of Add to Mailing List and Remove from Mailing List messages.

VMattie sent 7 reminder messages to seminar organizers on time during this testing. She received 5 replies to her reminders before the seminar announcement distribution date. She correctly inserted "TBA" for the remaining 2 instances in the seminar announcements.

Rigorous testing of CMattie is planned. CMattie's implementation of global workspace theory makes her an agent significantly more complex than VMattie. At the moment, even without the test results, CMattie's role as an implementation of global workspace theory makes her valuable as a conceptual model of mind.

8. Conclusions

This paper presents an overview of CMattie's architecture focusing on "consciousness" and conceptual learning. The two modules implementing these contribute to the implementation of global workspace theory, and allow her to interact intelligently with seminar organizers. This interaction succeeds due to the unique integration of these two modules. Perceptual output enters the focus, which is brought to "consciousness". The global workspace broadcast allows for "conscious" conceptual learning, completing the cycle. This cycle allows CMattie to acquire her socially situated intelligence. In particular, we hope to show that "conscious" software agents can be capable of essentially one-shot learning through interaction with a human. In future stages, extensions such as unlearning are planned.

CMattie is the first software agent designed to implement global workspace theory. As such, she can be considered the first "conscious" agent. It is hoped that the implementation decisions both provide testable hypotheses to neuroscientists and cognitive scientists, and that successful results will lead to more intelligent "conscious" agents.

9. Acknowledgements

The first author is supported in part by NSF grant SBR-9720314. The second author is supported in part by the Cancer Center Support CORE grant, P30 CA 21765 and by the American Lebanese Syrian Associated Charities (ALSAC). The third author is supported in part by NSF grant SBR-9720314 and by ONR grant N00014-98-1-0332. The authors also acknowledge the other members of the "Conscious" Software Research Group, including Art Graesser, Zhaohua Zhang, Aregahegn Negatu, Ashraf Anwar, Lee McCaulay, and Scott Dodson.

Notes

1. "Conscious" Mattie is intended to implement a psychological theory of consciousness, hence her name. This theory is described later in the introduction. We make no claims that she is conscious in the sense of being sentient.

References

Allen, James. 1995. *Natural language understanding.* Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc.

Anwar, Ashraf & Stan Franklin. Forthcoming. *Sparse distributed memory for "conscious" software agents*. http://www.msci.memphis.edu/~cmattie.

Baars, Bernard. 1988. *A cognitive theory of consciousness*. New York: Cambridge University Press.

Baars, Bernard. 1997. *In the theater of consciousness*. New York: Oxford University Press.

Bogner, Myles. 1998. *Creating a conscious agent*. Memphis: Master's thesis, The University of Memphis.

Boone, Gary. 1998. Concept features in Re:Agent, an intelligent email agent. *Proceedings of the Second International Conference on Autonomous Agents*. New York: ACM Press, 141–148.

Franklin, Stan. 1995. *Artificial minds.* Cambridge, MA: The MIT Press.

Franklin, Stan. 1997. Autonomous agents as embodied AI. *Cybernetics and Systems*, Special Issue on Epistemological Issues in Embodied AI, 28, 499–520.

Franklin, Stan & Arthur Graesser. 1996. Is it an agent, or just a program?: A taxonomy for autonomous agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures and Languages*. Berlin: Springer-Verlag, 21–35.

Franklin, Stan & Arthur Graesser. In Press. *Consciousness and Cognition.* http://www.msci.memphis.edu/~cmattie.

Hofstadter, Douglas & Melanie Mitchell. 1994. The copycat project: A model of mental fluidity and analogy-making. Holyoak, K. & Barden, J. (Eds.). *Advances in connectionist and neural computation theory, 2*. Norwood, NJ: Ablex.

Holland, J. H. 1986. A mathematical framework for studying learning in classifier systems. Farmer, D., et al. (Eds.). *Evolution, games and learning: Models for adaption in machine and nature*. Amsterdam: North-Holland.

Jackson, John. 1987. Idea for a mind. *SIGGART newsletter*, 101, 23–26.

Kanerva, Pentti. 1988. *Sparse distributed memory.* Cambridge, MA: MIT Press.

Kolodner, Janet. 1993. *Case-based reasoning*. Morgan Kaufmann Publishers.

Kozierok, Robyn and Pattie Maes. 1993. A learning interface agent for scheduling meetings. *Proceedings of the 1993 International Workshop on Intelligent User Interfaces.* Orlando, FL, 81–88.

Lashkari, Yezdi, Max Metral, & Pattie Maes. 1994. Collaborative interface agents. *Proceedings of AAAI '94 Conference*. Menlo Park, CA: AAAI Press.

Maes, Pattie. 1990. How to do the right thing. *Connection science.*

Maturana, H. R. 1975. The organization of the living: A theory of the living organization. *International Journal of Man-Machine Studies*, 7, 313–32.

Maturana, H. R. & F. Varela. 1980. *Autopoiesis and cognition: The realization of the living.* Dordrecht, Netherlands: Reidel.

McCauley, Lee & Stan Franklin. 1998. An architecture for emotion. *AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*. Menlo Park, CA: AAAI Press, 122–128.

Mitchell, Melanie. 1993. *Analogy-making as perception*. Cambridge, MA: MIT Press.

Ramamurthy, Uma, Stan Franklin, & Aregahegn Negatu. 1998. Learning concepts in software agents. From Animals to Animat 5: *Proceedings of The Fifth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press, 372–377.

Ramamurthy, Uma, Myles Bogner, & Stan Franklin. 1998. Conscious learning in an adaptive software agent. *Proceedings of The Second Asia Pacific Conference on Simulated Evolution and Learning (SEAL98)*. Canberra, Australia.

Sycara, Katia & D. Zeng. 1994. Visitor-hoster: Towards an intelligent electronic secretary. *CIKM94 Workshop on Intelligent Information Agents*. http://www.cs.cmu.edu/afs/cs/user/katia/www/visit-host.html.

Song, Hongjun. 1998. *Control structures for software agents*. Memphis: Doctoral dissertation, The University of Memphis.

Wheeler, Michael. 1994. *For whom the bell tolls? The roles of representation and computation in the study of situated agents*. Cognitive Science Research Paper 320, School of Cognitive and Computing Sciences, University of Sussex.

Zhang, Zhaohua, Stan Franklin, & Dipankar Dasgupta. 1998. Metacognition in software agents using classifer systems. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*. Menlo Park, CA: AAAI Press, 83–88.

Zhang, Zhaohua, Stan Franklin, Brent Olde, Yun Wan, & Arthur Graesser. 1998. Natural language sensing for autonomous agents. *Proceedings of the IEEE Joint Symposia on Intelligence and Systems*. Rockville, Maryland, 374–81.

CHAPTER SIX

Emotionally Grounded Social Interaction[1]

Dolores Cañamero and Walter Van de Velde

1. Introduction

The question we address is the following: how do agents that are embedded in large societies — hundreds to millions of agents — focus the enormous potential for interaction with other agents towards the most productive encounters?

The relevance of this question needs little argument. From a social studies viewpoint, it is tied to the understanding of a society's structure and how it — implicitly and explicitly, directly and through internalisation mechanisms — regulates daily interaction. From a multi-agent (MA) studies perspective, the issue is fundamental to the understanding of artificial societies. In particular, it enlarges the scope of behaviour control in agents, which is not only aimed at task achievement (e.g., finding food to survive), but at doing so in an evolving social context that, by the interdependencies it entails, makes every strict assumption of self-sufficiency a naive one. More technically, one cannot expect exhaustive exploration of potential encounters to be an efficient means of implementing large-scale agent societies. This is especially true for open MA systems — those in which the agents are being designed independently and without a predefined and globally known relationship towards the others. Finally, and here lies another important motivation for our work, such large-scale open agent societies will soon be of practical relevance — we predict that networks like WWW will evolve into large collections of software agents independently pursuing goals on behalf of their users, or trying to offer services to users and other agents.

In pursuit of an answer this paper brings together two lines of research. One line has focused on understanding large heterogeneous societies in terms of separate but loosely coupled communities, and on developing technologies that could enhance social processes within such societies. This work on *co-habited mixed reality* (Van de Velde 1997) draws inspiration from anthropology, in particular concerning the role of bodily appearance and presentations to encode relationships of social stature, power and representation (see for instance Keane 1997). It studies the processes that result in the structural coupling between the different constituent communities of a complex society. The second line of research that feeds into this paper is on bodily-grounded emotions. This work stresses the role of affects as mechanisms for adaptation, both from an evolutionary perspective and in the sense that this term has in (Ashby 1952) — contributing to maintain the organism (a dynamic system) viable in its environment, i.e., internally stable. Therefore, emotional states are not directly modelled at a high, psychological level but grounded in a physiology that ensures appropriate feedback between behaviour and affective states in the interactions of the agent with the world. This work models an autonomous agent in terms of a synthetic physiology which grounds concepts like motivations, drives, and emotional states. These are linked in an internal dynamics for regulating physiological parameters through behaviour selection that, phenomenologically, leads to intentional and adaptive behaviour of the agent (Cañamero 1997a). This work aims to understand and exploit some of the roles that emotions play in biological systems in order to develop mechanisms and tools to enhance autonomy, adaptation, and social interaction in artificial societies.

Both research lines attack the same underlying problem: understanding the mechanisms that regulate individual behaviour and social interaction in large open agent societies. The work on grounding emotions had previously focused on the impact of affective phenomena on individual behaviour, but had not yet exploited their external manifestations for purposes of social interaction. The work on cohabited mixed reality tried to give a new role in software agent research to bodily expression and its perception. However, it has approached this from an intentional level of *interest models*, a direction which in the long run does not seem sustainable (the origins of interest and especially their communication are problematic). The combination of both approaches fleshes out — forgive the wording — elements of a functional view on the body in the context of multi-agent research and artificial life. To

avoid all misunderstanding: our work is not on physical embodiment, neither is it related to research on avatars or virtual presence. We want to understand the various roles of physical embodiment, rather than proclaiming its necessity for realising true intelligence. On the contrary, we claim that such roles remain valid and useful in a non-physical or hybrid setting. Our work also has an applied dimension which is pursued in an ongoing European project, called COMRIS.[2] The aim of this project is to support participants in information-intensive events where the possibilities for interaction are many, but not every encounter leads to a productive result. For example, at a major fair there may be many potential buyers and sellers, but only some of them will come to a productive interaction — a business deal. Our technology aims to focus this large potential for interaction so that the effectiveness of participation to such events is enhanced. The results in this paper are preliminary. We have not arrived yet at a full experimental integration of our work, nor do we have a final answer as to what a good experiment would look like. Especially when involving a mixed human and software agent setting, as is the case in the COMRIS type of application, quantifiable success criteria remain hard to pin down. However, the conceptual framework that results from combining our approaches provides a setting for attacking a variety of aspects of the problem of social intelligence.

2. Co-habited Mixed Realities

At the heart of an approach to supporting social processes must be a hypothesis on the structure and functioning of societies. Rather than with the modelling of specific processes (negotiation, work flow, coordination) our concern is with capturing what we could call the social engine, i.e., the process that underlies *any* manifestation of social interaction. Rule-based inference is a similar engine, this one for reasoning, that has been claimed to underlay any manifestation of individual intelligence (Newell and Simon 1972). A social engine should not be confused with social convention: our intended meaning is that the rules of a social engine are the nature of any social entity, i.e., can not be broken, just like a rule-based system has no choice but to follow the rules. Also note that hypothesising a unique and universal social engine is a dangerous claim — see what happened to rule-based inference — but we take it for now as a valid working hypothesis. Below are four conceptual building blocks for our approach to this.

Agency

The degree of agency of an entity can be assessed along three dimensions: autonomy, awareness, and historicity. Note that we do not define 'agent' as such, but provide a framework to assign a degree, qualitative for now, of agency to an entity. We use the term *agent* in the rest of this paper to refer to any entity that has a non-trivial degree of agency. At minimum, agency requires a potential for behaviour, that is, an observable interaction between an entity and its environment. Our notion of environment is different from the catch-all characterisation of everything-except-the-agent. We view an environment as a collection of resources. For convenience this collection is normally spacio-temporally organised. However, since space and time are resources themselves, the organisation along some primary resources is probably not essential, but relative to certain communities living in them. From this notion of environment then follows a definition of interaction as exchange of resources between an agent and its environment. Such exchange is not necessarily reversible or resource-preserving, and it may have direct effect on other resources. Unlike others we do not see the agent-environment interaction as essentially communicative (Genesereth and Ketchpel 1994).

- *Autonomy* refers to the extent in which an agent's potential behaviours exhaust those that are allowed within the general framework of a social engine. An autonomous agent exploits its freedom.
- *Awareness* refers to the extent in which an agent is embedded in its environment, in such a way that the resulting behaviour is in some sense better adapted to it. From an observer's point of view, an aware agent is perceptive to the changes in its environment and to the impact of its behaviour on it (to be understood, again, in terms of resources).
- *Historicity* refers to the extent that the entity has a history, i.e., that its future behaviours are conditioned by its past interactions. This idea is closely related to Dautenhahn's notion of autobiographic agents (Dautenhahn 1996). A capacity of learning is normally understood here, but this term is too restrictive as we want to encompass forms of adaptation and structural coupling.

Cohabited mixed reality

A cohabited mixed reality is a decomposition (not necessarily a partition) of a complex society (agents and environment) into a set of loosely coupled *communities*. A community is an open collection of agents that share a

resource, i.e., have the same or a related interest in it. A community is contrasted with a group in which the binding factor is a task or goal. The nascent field of communityware, the technology for supporting communities, therefore, clearly has a different aim than computer supported cooperative work (CSCW). The coupling between two communities consists of distinct points in which two members of different communities share part of their perceived environment (i.e., their awareness).

Interest-based navigation

Interest-based navigation is the process by which agents in a society seek contact with those other agents with which an interaction is most likely to further its interests. This process, in other words, is the process dual of our structural notion of community. It biases agents to interact with the members of their evolving communities. Obviously an agent may be part of several communities and, as stated before, a community is not a fixed and closed collection of agents.

Competition for attention

Competition for attention is the process by which agents engage in actual interaction with each other, 'negotiating' on the selection of behaviours that are likely to further their mutual interest. It is the process that focuses attention of agents on those series of interactions that are most likely to lead to a productive encounter for all parties involved. Although attention is a resource itself, one that can be exchanged, traded, stocked and invested in, we give it a special status because of our purpose of explaining social processes. At the same time we want to push this metaphorical idea (and the others just described) to a level of concretion where it becomes a source of technical inspiration.

It would lead us too far to spell out all the motivations and details of this framework. Instead we will focus on the type of cohabited mixed reality that the COMRIS project is aiming at. In COMRIS, the co-habited mixed reality consists of a pair of a real and a virtual space (Figure 1). The real world is inhabited by humans, whereas the virtual one is inhabited by software agents. Note that we keep both worlds apart. Therefore, and contrary to augmented-reality models, we are relieved from the task of establishing a mapping between both realities, as well as from having to try any kind of perceptual integration. For instance, in real space the structure (topology) is based on

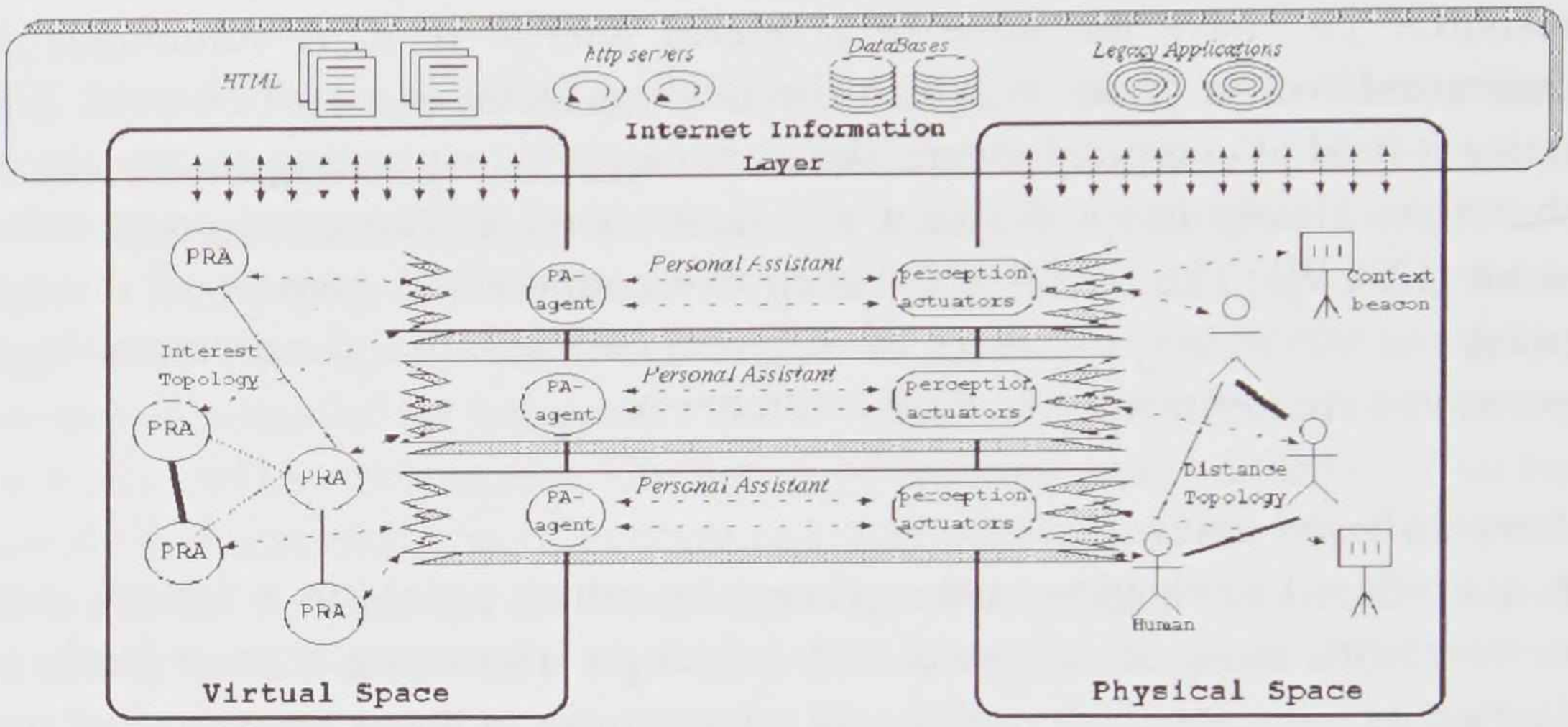


Figure 1. The co-habited mixed reality (COMRIS) technical framework

physical distance. In virtual space this can be different, as in our case, where distance is a measure of interests. This different topology reflects the fact that the environments, constituted as different resources that are handled and the way in which the agents interact with them, are different in both spaces. As far as the inhabitants of one space are concerned, the other world is not even directly perceivable, neither do they have a feeling of virtual presence, nor are they represented in it by some sort of avatars. We only need a minimal awareness (the loose coupling) from one space to the other, so that processes in the real world can enhance the ones in the other.

A good example of this type of co-habited mixed reality is provided by electronic commerce. The obvious meaning of an electronic market place is that it is a market place for us, humans, that happens to be realised in a digital medium. We are browser-shopping in on-line stores, exploring digital catalogues, filling our shopping cart with goods. With the introduction of software agents we get help from digital assistants, and we can for example delegate the task of searching for a good deal to them. However, as more and more commercial roles are being supported by software agents, the role of the humans in it becomes marginal. The electronic market becomes a parallel market, in which software agents are doing their businesses. They buy and sell, they advertise, they build up long term relationships, alliances and markets. It just so happens that some of them are driven by the needs in the real world. It just so happens, also, that some of the goods they are trading have a

manifestation in the physical world, but they do not need to care, except for sharing an awareness of the demands and offerings of some good. They could even use a totally different type of currency.

The link between the different communities in a cohabited mixed reality can take many forms. In the COMRIS case it is a wearable electronic device that each real participant wears as her Personal Assistant (PA in Figure 1) — the COMRIS *parrot*. The COMRIS parrot (also called by its species name *psitacus comris*) is an example of a wearable computer (ISWC 1997), a computing device that is worn by the user much in the sense that cloths or shoes are worn. This idea of the COMRIS parrot is inspired by the image of a parrot sitting on one's shoulder, looking around with you and whispering relevant information in your ear. Like any good parrot, at least if it belongs to the mythical sailor, it has this extra sense, as if it lives in another world that allows it to perceive things which you cannot (like an upcoming storm). A wearable computer lives with its host, and may serve a variety of purposes on an ongoing and situation-sensitive basis. Situated information access and automated memorisation and remembrance are typical examples. In COMRIS we want to explore it for enhancing social processes.

The COMRIS parrot is an original device in many ways. It is a full audio device using speech output, as opposed to others that use screen-like output, possibly projected on the retina or displayed in a normal pair of glasses. A notable exception is (Sawhney & Schmandt 1998). Another distinguishing feature is that it is a link into a parallel society, rather than just being an extension of the wearer's memory or information space. The parrot is currently under development at Riverland Next Generation and it is based on a StrongARM processor (the first prototypes will be running after summer 1998.) It is able to locate the host with respect to the space he or she is in. No exact positioning is required, but presence in a room or closeness to some other must be detected. The device is also able to locate and identify devices worn by other participants, and identify them. Finally it communicates wirelessly with an Intranet host where the main cognitive part of the personal assistant is running. Future versions will be able to also infer the activities of the host. This will result in a better awareness or *context perception*, and thus of a better coupling between the spaces. Note that this device, like a true wearable, requires no explicit input from the so-called user (you do not use your cloths, really). Its output is the result of the process of competition for attention between the Personal Representative Agents (PRA in Figure 1) in the virtual

space. This pointing to relevant information falls into the *radical information push* philosophy of COMRIS. In this model information is actively imposed upon the user in its concrete minute-to-minute context of activities. The trend toward information push is a consequence of increasing agentification of the Internet, and incidentally also the original motivation behind the ideas of competition for attention (Van de Velde et al. 1997). The widely held view that intelligent agents are (just) tools to manage the complexity and size of on-line information is narrow minded. Although search for and retrieval of information is an important activity on the net, it hides the underlying purpose. In the near future computer-networks will be populated with thousands of software agents actively trying to push their interests in competition with each other for the attention of the user. The effectiveness of this idea depends on (1) the capability to detect the context, situation, and interests of the host on a minute basis (awareness), (2) the capacity of the virtual space, its inhabitants and the processes going on in it to focus on activities that are relevant for the user's interests (interest-based navigation), and (3) the capability to render, in a user friendly way, the information landscape as it relates to the user's activities.

3. Emotions

Emotions are complex, multifaceted phenomena, the nature of which cannot be grasped by a single definition. As far as the layperson is concerned, emotions are primarily subjective experiences or feelings which lie at the heart of what being human is all about. From a more scientific standpoint, emotions encompass neuro-endocrine, physiological, cognitive, and social aspects of behaviour. Their study has been undertaken by different disciplines, such as psychology, neurobiology, and social and cultural studies, and different aspects have been stressed by diverse research paradigms (see Cornelius 1996, for an overview). However, underlying the different theories there is a common insight: emotions, whatever interpretation is given to them, *serve a purpose*. We adhere to this functional view of emotions, as Frijda (1995) calls it.

3.1 *Emotions and other affective phenomena*

The type of phenomena that are included under the category of emotions varies, and the boundaries with respect to other affective states are not always

crisp. We distinguish between motivations and emotions (Cañamero 1997a), and assign them different roles in behaviour selection and adaptation. Motivational states, such as hunger, thirst, aggression, etc. are drives toward action based on internal bodily needs related to self-sufficiency and survival. In general, motivations can be seen as homeostatic processes which maintain a controlled physiological variable within a certain range. Therefore, they involve arousal and satiation, and vary as a function of deprivation. According to (Kandel et al. 1995) and in our model, motivations have three main functions: (1) they steer behaviour toward, or away from, a specific goal; (2) they increase general alertness and energise the individual to action; and (3) they combine individual behavioural components into a goal-oriented behavioural sequence.

Emotions have to be integrated in artificial systems to promote decision making, activity selection, and autonomy. The homeostatic model, although fine for explaining motivational mechanisms, is less adequate in this case, in particular for the notions of arousal and of activation as its indicator. Contrary to activation theories, which state a mere correlation between the amount of hormone released, amount of neural excitation, and amount of emotional arousal, we follow Pribram (1984) to view emotional activation as an indicator of a change in configuration of neural and endocrine activity with respect to the habitual stable baseline of the organism. One can thus view emotions as second-order modifiers or amplifiers of motivation. Another fundamental difference between motivational and emotional states is that, whereas motivations are highly urgent and specific, emotions combine both, structural/functional specificity and generality of object, time, intensity, and density (Tomkins 1984). But if emotions show this generality, what accounts for the activation of different affects? Tomkins proposes three variants of a single principle: stimulation increase, which activates both positive and negative affects such as startle, fear, and interest; (high) stimulation level, which only activates negative emotions such as distress or anger; and stimulation decrease, which only activates positive emotions such as joy.

3.2 *Emotions and the individual*

Emotional states are adaptive mechanisms that allow to deal with events, both internal and external, which are important for the survival of the agent (and the species) in a particular environment (LeDoux 1996). In this sense, they are considered as remnants of the creature's evolutionary past. For example, the

perception of a potentially life-threatening event, such as the presence of a predator, immediately puts the animal in a fear state; this state acts as a defence mechanism by eliciting a fight-or-flight reaction that prepares the organism to respond rapidly to the external environment (for example, a transient, rapid increase in blood pressure permits the animal to run faster).

Emotions play a major role in motivating and guiding action. At the simplest level, the categorisation of events as pleasant/unpleasant, beneficial/noxious, turns neutral stimuli into something to be pursued or avoided. This idea lies at the heart of operant conditioning theories of learning. Due to their generality of object, time, intensity, and density, emotional states amplify the effects of motivations, which are stimulus-specific and need an urgent satisfaction (Tomkins 1984). The role that emotional systems play in decision making (activity selection) and social relations has been evidenced by Damasio's studies of patients with damage in the prefrontal cortex and the amygdala (Damasio 1994). When confronted with situations requiring to choose between alternative options, these patients are perfectly capable to analyse — over-analyse, indeed — the consequences, pros and cons of each possible option, but they are unable to make a decision so as to what to do; abnormal social relations and inability to properly assess social situations are also observed in these patients.

3.3 *Social character of emotions*

Emotions come into play mainly (but not only) when we consider an individual in interaction with its social environment. Emotions mostly arise and make sense in a social context, and they are shaped by it to a great extent. Let us examine these two issues a bit more in detail.

The external manifestations of emotions can play a major role as signalling mechanisms (although this does not imply that emotions evolved primarily as a communication tool) at several levels. First, they play a role of social reference (Dantzer 1994). The emotional expression of an individual can be used by another to assess the type of situation it is confronted with and adapt its behaviour accordingly. In some cases emotional expression seems to have a communicative role. Although the external manifestations of emotions can be seen as merely signalling the achievement of some survival-related goal, they can also be controlled to some extent and be intentionally used to let the others know one's emotional state so that these can make expectations about

the situation they are confronted with, and thus of the best behaviour to follow. An example from the animal world is the case of the diverse alarm calls emitted by monkeys in different fear-related situations and that elicit different behaviours in conspecifics (see for instance Kalin 1997). Emotions can thus be said to carry some information content.

Emotional states also contribute to the construction of intersubjectivity. For example, an animal will rarely attack another that shows signs of submission or sadness. In this sense, emotions contribute to the elaboration of the image an individual has of itself *with respect to the partner* it is interacting with. This self-image is a reflection of the image we have of the other, and elicits the mirror process in it, which in turn can elicit an emotional/behavioural reaction on my side that shows my agreement on the image the other has constructed about me, giving thus rise to the construction of a negotiated intersubjectivity.

Turning to the origins of emotion, it is unclear to what extent the expression of emotions is a product of nature, and to what extent it is influenced by culture in the form of display rules. Social constructivists, such as Averill (1980), have made special emphasis on the socially constructed nature of emotions to consider them as a special kind of social role. For these authors, emotions are not characterised by a set of particular physiological features that must occur for a certain emotion to be present, but by a socially prescribed and learned set of responses to be followed by a person in a given situation. Emotions also carry social values and are to a large extent moral judgements. They may thus fulfil a social function. Fear, for example, can be seen as instrumental in sustaining social values (Armon-Jones 1986); it certainly helps to keep us from harm, but also contributes to keep us in our place in society, as individuals are usually taught to be afraid of things that imply a violation of what the society expects from them and tolerates.

The way in which we experience and categorise emotions is also largely influenced by society. In Western cultures, emotions are categorised according to self-centred internal states or subjective feelings, and experienced as passions that take over our mind and behaviour and over which we have very little or no control. This way, an individual acting under the influence of an emotional state is in general waived from all responsibility, since it can be said that he is not being himself at that moment. On the contrary, in other cultures such as the Ifaluk, emotions seem to be categorised according to the social situations in which they occur, as reported by Lutz (1988).

The constructivist nature of emotions is, however, not in contradiction with the fact of emotions corresponding to internal states, even though they can be defined and elicited by social relations. Making this link is the objective of the next section.

4. Agency and Emotion

4.1 *The COMRIS agent model*

The model of a COMRIS agent, as it was proposed in (Van de Velde 1997) is depicted in Figure 2. This model prescribes a conceptual interface, and does hardly imply an architecture. It is, however, deliberately simple and minimal to allow for the easy addition of agents, and the interaction between different agents without necessarily sharing extensive domain ontologies. The latter is, after all, only required if agents decide to engage in a domain specific interaction, e.g., a commercial transaction or any other collaborative activity. what interests us most here are the processes that precede this.

We start out by integrating the COMRIS agent model with the agent architecture. A central notion in COMRIS is the one of interest: agents are conceptualised as having an interest, the pursuit of which drives social interactions. The social engine that COMRIS proposes consists of the processes of competition for attention and of interest-based navigation. Techniques of *interest-based navigation* bring together those virtual agents whose interests are likely to merge into a productive social process. To deal with the complexity of finding out about useful partners we propose the notions of *appearance* — the external bodily features that partly reflect the agent's internal state — and of *presentation* — the set of external features that the agent shows to the agents it interacts with, depending on its interests in the communication process, its expectations about the others, etc. On the other hand, different agents that have converged toward interacting with another one, may have to compete to get this other's attention at a given point. At all times, *competition for attention* mechanisms (Van de Velde et al. 1997) that take into account several factors contributing to define the internal state of the agent, such as its competence, its performance, or the relevance of its action in a particular information context, drives the interactions among the agents.

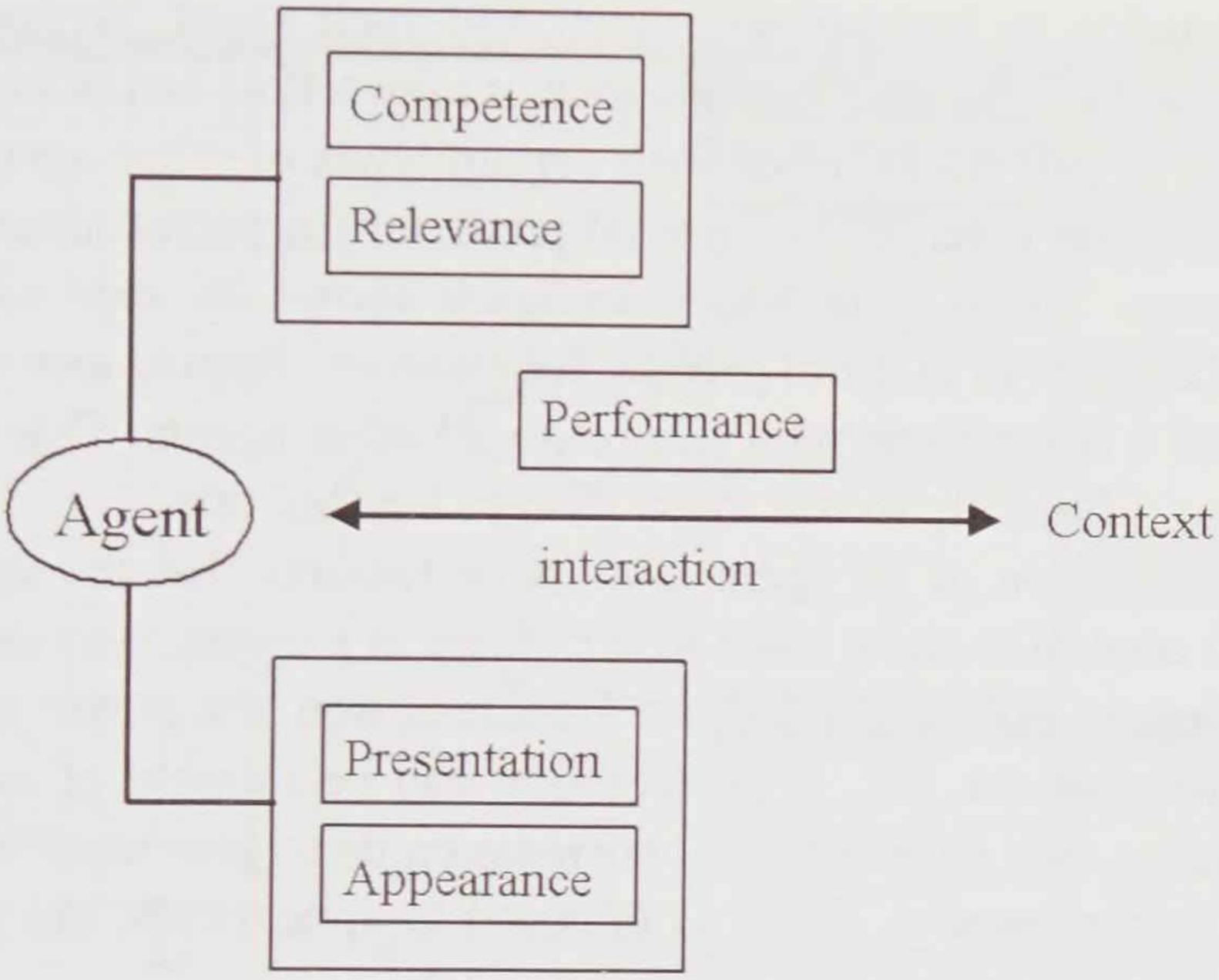


Figure 2. The COMRIS agent model

The model components that are important for competition for attention are the competence of an agent's behaviour, its relevance, and the performance of the agent. *Competence* refers to the extent to which an agent believes it is good in doing a behaviour of which it believes it furthers its interest. *Relevance* is the adequacy of such behaviour with respect to the present situation. The agent's *performance* reflects the degree of success in achieving its interests in its interactions with the others. Competence and relevance are communicated to the other agents for whose attention the agent is competing. In COMRIS this communication is in numerical terms. However this cannot be a final solution. Instead, and in relation to the agent architecture, we now see them as triggering certain behaviours, and thus as motivations or drives. The fact that in this case the success of the behaviour requires a response of another agent, implies that these behaviours will be essentially communicative. Emotional modifiers will reflect the temporal evolution of level of satisfaction of an interest, where this satisfaction level can be directly tied into the agent's physiological process.

The social attitude (or embedding) of an agent is expressed to the outside — to the others — through a twofold mechanism of appearance and presentation. The *appearance* of an agent is a set of external bodily features that partly express the way in which an agent positions itself within a multi-agent society,

i.e., its inclination to interact with such and such agent. Appearance may reflect historical information. For instance, it may reflect implicitly the results (good or bad) of past interactions with certain types of other agents. Another way to put it is that it serves as a partial predictor for performance values for certain interests. The way in which an agent shows its appearance can be linked to social roles or cultural groups, for instance. Appearance may change over time, but it is the same with respect to all other agents. That is, an agent cannot appear differently to one agent than to another one.

The *presentation* of an agent is a set of features that the agent can (to some extent) control to show itself to the others in a certain way depending on its interests in and expectations about the interaction, the power relationships with the other partners, etc. A presentation can be thought of as a business card. Each agent may have different business cards to give to different people in different circumstances. Thus, as opposed to appearance, the presentation of an agent can be targeted and adapted as such.

4.2 *Emotions in viability and adaptation*

The work on emotions in (Cañamero 1997a) has proposed a genuine agent architecture. Here, an agent has a synthetic physiology that defines its internal state and that makes it have motivations driven by internal needs, try to execute behaviours that contribute to satisfy those needs, and enter in emotional states as a consequence of the effects that its interactions with the world have in its physiology and that feed back in it and in the agent's behaviour. The experimental setting in (Cañamero 1997a) is a dynamic two-dimensional world called Gridland (Figure 3) and inhabited by two species — the smart agents (As) and the enemies (Es) — and which also contains food and water sources and blocks — geometrical figures — of varying shapes and sizes.

The creatures are implemented as a collection of modules of different types. *Es* have a simple, rather stereotyped behaviour, their main role being to introduce more dynamicity in the world — they wander around avoiding obstacles, try to ingest every form of organic matter they run into, and withdraw whenever they feel pain, caused by another creature's attack. *As* are more complex creatures used as a testbed for affects. Among other elements, their architecture includes: internal and external sensors, maps, behaviours, both appetitive (go-toward, look-for, find) and consumatory (eat, drink, play, rest, withdraw, etc.), motivations (e.g. fatigue, thirst, boredom) and basic emotions (fear, anger, happiness, etc.).

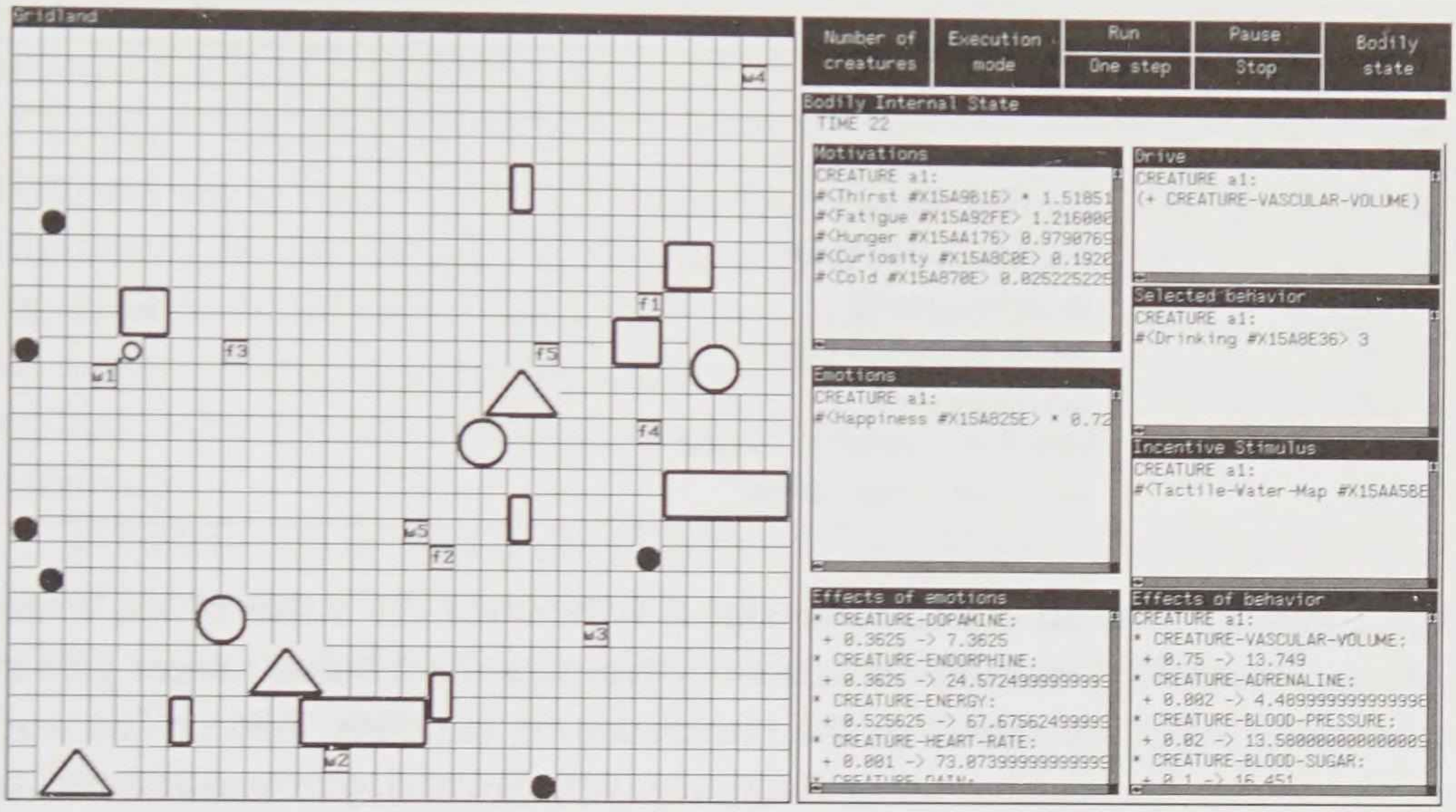


Figure 3. A snapshot of Gridland. As are empty 'dots' with a black spot in one corner — their eye. Full dots are Es

In addition, *As* have a *synthetic physiology* — a set of parameters that define their internal bodily state and needs (see Table 1). This physiology includes both, controlled variables necessary for survival (e.g. heart-rate, energy, blood sugar level) and hormones released under different emotional states that modify the amount of the controlled variables. Controlled variables are monitored by motivations (see Table 2) implemented following a homeostatic model.

Table 1. Physiological variables used to define the agents' (As) bodily state in Gridland

Parameter	*Initial value*	*Set-point*	*Variability*
Adrenaline	10	10	± 5
Blood pressure	12	12	± 4
Blood sugar	30	20	± 10
Dopamine	10	10	± 5
Endorphin	20	20	± 10
Energy	120	100	± 50
Heart rate	75	75	± 25
Pain	0	0	± 2
Respiration rate	8	8	± 7
Temperature	37	37	± 3
Vascular volume	25	20	± 10

Table 2. As' motivations and their drives

Motivation	*Drive*
Aggression	decrease adrenaline
Cold	increase temperature
Curiosity	increase endorphin
Fatigue	increase energy
Hunger	increase blood sugar
Self-protection	decrease pain
Thirst	increase vascular volume
Warmth	decrease temperature

Each motivation receiving an error signal (whenever the value of the corresponding controlled variable moves away from its ideal value range) gets an activation level proportional to the magnitude of the error, and an intensity calculated on the basis of its activation level. Several motivations can be active at the same time, but that with the highest activation level gets the creature's attention and tries to organise its behaviour so as to satisfy its main drive. The strongest motivation selects the (consumatory or appetitive) behaviour(s) that can best contribute to the satisfaction of the most urgent need. Table 3 shows the repertoire of consumatory behaviours implemented in the agents, the incentive stimulus that must be present for a behaviour to be executed, and the main effects that the execution of the behaviour has on the agent's physiology.

Table 3. As' repertoire of behaviours, their incentive stimuli, and their main effects on the agent's physiology

Behaviour	*Stimulus*	*Main effect*
Attack	living-being	decrease adrenaline
Drink	water	increase vascular volume
Eat	food	increase blood sugar
Play	block, A	increase endorphin
Rest	top of flat block	increase energy
Walk avoiding obstacles	free space	increase temperature
Withdraw	pain	decrease pain

A second motivation is also taken into account to allow for opportunistic behaviours. This way, a creature searching for food as its main motivation, but that is also thirsty, will stop to drink if a water source appears on its way, and will continue its search for food afterwards. However, it will not actively look

for water unless the priority of its motivations changes. The execution of the selected behaviour modifies the values of different controlled variables with an intensity proportional to that of the motivation, bringing the variable that triggered the drive to its ideal value, and therefore satisfying the need that originated the behaviour. Motivations thus drive behaviour selection and organisation based on the notions of arousal and satiation, and also determine the agent's focus of attention. They can be thought of as implementing an implicit value system that ensures decision making, activity selection, and autonomy.

Emotions are second-order modifiers or amplifiers of motives and therefore of behaviour. Table 4 shows the set of basic emotions implemented in *As* and the events that trigger them. Emotional states are signs of internal instability, but they also seem to play a role in homeostasis as mechanisms of re-equilibration, as processes that try to stop ongoing behaviour, complementing this way the role of motivations — appetitive processes that try to activate action as a response to deprivation. In our model, emotions are implemented following a hormonal model (Cañamero 1997b). They exert further control of the agents behaviour by sending synthetic hormones that may affect not only the agent's controlled variables, but also its perceptual, attentional, and motivational mechanisms, this way modifying the intensity and execution of the selected behaviour.

Table 4. Innate stimuli triggering As' emotions

Emotion	**Triggering event**
Fear	presence of enemy
Anger	accomplishment of goal menaced or undone
Happiness	achievement of goal
Sadness	inability to achieve goal
Boredom	repetitive activity
Interest	presence of novel object or event

5. Social Interaction and Emotion

In our view of sociality that underlies the expression 'social interaction' in this paper, we follow Maturana and Varela (1987) to see social phenomena as the *structural coupling* of agents with their environment and among themselves, where structural coupling is the history of recurrent interactions leading to the

structural congruence between two or more systems. This is particularly clear in our notion of cohabited mixed reality, where the different communities can be considered to be structurally coupled. In structural coupling we are looking for ways in which agent societies acquire mechanisms for coordination. We are not trying to understand rational society models, i.e., those that are based on careful top-down organisation for optimising some work organisation or other process. This we leave to workers in computer supported cooperative work. Instead, we are interested in understanding how social communities are being structured by their elements, using implicit rather than explicit mechanisms. In particular a variety of elements, physical objects and emotional expressions among others, are being used to impose a structure on the interactions that are taking place in a complex society.

The integration of the original model in the COMRIS agents requires some modifications and extensions to accommodate the demands of complex social interactions, as well as those of an information context. Let us consider the main ones. First, the physiology of the agents must be augmented to include parameters which are relevant for an agent that has to survive in an information context: parameters to keep track of the amount and relevance of information, of the agent's performance, of its competence, etc. Motivations, implemented again according to a homeostatic model, monitor these variables by selecting behaviours whose execution modifies the parameter in the desired direction. In the same way, the particular modules integrating the agents (sensors, maps, behaviours, etc.) must be adapted to this particular world. As for emotions, we propose to complement the set of basic emotions used in the original model, each implementing an adaptive function fundamental for the survival of an agent in its particular environment, with a set of social emotions that come to play only in the context of social interactions, and that should contribute to the adaptation of the agents to their social environment. Therefore, the main extensions have then to do with the ability to show, perceive, and react to emotional expression, as we will see in the next section.

The concrete direction that we explore here is thus to use *bodily-grounded emotions* and their various expressions as an essential part of appearance, presentation, and the emergence of long-term co-behaviors. In the original COMRIS model the agent features of appearance, presentation, competence and relevance are used to regulate interest-based navigation and competition for attention. Having adapted the agent's emotional apparatus, we can take the crucial step of involving *emotional expression* in each of these factors. Our

software agents are also endowed with other affective phenomena, equally rooted in their physiological state, such as survival-related motivations driving them to act autonomously, moods, and temperaments that contribute to define their individual personalities. We explore at least two ways in which emotional expression is relevant for either one of the basic processes in COMRIS.

5.1 *Social emotions*

The role of emotions in the COMRIS model can be sketched out as follows. We remain faithful to the view emotions are second-order modifiers of motivational processes. These latter are captured by the notion of interest in COMRIS which, in a particular context, leads to a behavioural tendency. Emotions work as amplifiers or amortisers of motivational behaviour, contributing in turn to the regulation of the agent's physiology. Emotions are indicators of unusual progress and not of the normalisation process. The reason then why such emotions are not stimulus-specific is because a direct link to the particular behaviour would not help since its problem cannot be solved within the closed world of that behaviour alone.

Emotional expression, and this is the link with the model components of appearance and presentation, comes into play when an agent's behavioural repertoire is insufficient to resolve the emotional state. In this case, the agent needs to take resort to others. It will express its emotional state to signal for help to the other agents. These, by their interaction with the agent or its environment are hoped to be of use to the agent, thus indirectly resolving its emotional state. Should this turn out not to work, the other agents may end up being in the same or a similar emotional state. In this way emotional tendencies may spread through communities.

5.2 *Negotiating difference: Emotional rhetoric*

The model of social emotions that we propose here is inspired by that of Aristotle in his *Rhetoric*, which considers emotional states as tools to negotiate social interactions. This view starts by accepting the fact that individuals are different, and that they have to live together; in order to do so in harmony, they must take into account their differences and accept them or negotiate them. Emotions are signs of a difference between individuals, and at the same time the tools that allow them to negotiate this difference, fulfilling various roles —

epistemic, rhetoric, ethical, and political. The model of affects that Aristotle proposes in its *Rhetoric* is based on a couple of relationships between individuals: identity/distance (according to their interests, affinities) and symmetry/asymmetry (rank, status), as perceived by each partner. A social interaction between two individuals will produce an emotional state in each of them according to its own perception of the relation along these dimensions and of the image that the other shows about it. The effect of an emotional state, as far as social interaction is concerned, will be to try to maintain, increase or decrease the distance in each dimension, through external manifestation such as emotional expression and/or the execution of an appropriate behaviour. For example, love between A and B implies the mutual perception of an almost-identity (empathy) and of symmetry that both individuals will try to maintain or increase. Emotions can also activate each other and form rhetoric chains to negotiate distance and symmetry. For example, an overconfident individual A engaged in an interaction with B will perceive itself as having a much higher rank than B, disregard the distance dimension, and disdain whatever B proposes or does towards A. If the perception that B has of this relation is that A overestimates its self-image and underestimates B (i.e., the relation is not as asymmetrical as A seems to believe), B will enter in an angry state that will lead it to do something that increases the distance and decreases the asymmetry as perceived by A ("Hey, who do you think you are talking to?!!"). Emotions form a sort of interactive network, with opposition (e.g., love versus hate) and inhibition relations (e.g., anger and fear cannot be active at the same time).

In contemporary sociology, Kemper (1978; Kemper and Collins 1990) has proposed a rather similar model to analyse social relations and their emotional outcomes. Kemper takes into account two dimensions, power and status, to express social relations and emotions, which are seen as responses to the power and status meanings of situations. Power can be said to be based on symmetry relations, since it is considered as "a relational condition in which one actor actually or potentially compels another actor to do something he or she does not wish to do." (Kemper 1993: 42) On the other hand, status seems to be grounded on a distance dimension, as it is "the relational condition of voluntary compliance with the wishes, interests, and desires of another person." (Kemper 1993: 42) According to Kemper, a large class of human emotions result from real, anticipated, imagined, or recollected outcomes of social relations. From every individual's perspective, the outcome of a social interaction can be an increase, decrease, or no change in the self's power or

status with respect to the other, and an increase, decrease, or no change in the other's power and status with respect to the self. The particular emotions elicited by the 12 possible outcomes resulting from combinations of these factors will depend not only on the effects of the interaction on the power and status of each partner, but also on the agent who is considered as being responsible for the outcome — self, other, or a third party.

5.3 *Emotional expression and recognition*

Emotional expression can be kept simple, and however rich enough to meet the requirements of both appearance and presentation. Each emotional state automatically carries some external features over which the agent has no control (for example, a colour associated with each emotion) and that affect its appearance — a pattern of features or a number that reflects the agent's self-perceived rank or competence regardless of other individuals. This rank or competence level can be a function, for instance, of the history of successful interactions (e.g., those that increased the 'pleasure' variable above a certain threshold), and the amount and relevance of accumulated information or wealth. Ranks can belong to a culturally determined typology, and establish hierarchical relations in the society. The rank of an agent then tells at least as much about its social type (to which a social role can be associated) as it does about its history (how successful it has been in its interactions); it is thus a *social identifier*. In addition, the agents should be able to control some of the components of emotional expression so that they can choose how to present themselves to the others depending on their state, interests, and the image they have of the other, i.e., of their view of the distance and symmetry dimensions of the relation. Presentation can be expressed by modifying the pattern or number used to show appearance as a function of the social emotional state. In this way, presentation reflects the agent's self-image in the context of a social relation, reflecting also the image it has of its partner. This image of self and the other has also a behavioural impact in the interaction — for example, in the price that an agent proposes to exchange goods or information with another.

As for emotion recognition, we propose the use of expressive patterns characterising the different emotions. In particular, the difference between the appearance and presentation patterns of an agent *A* can be used by another agent *B* to recognise the image that *A* has of the relation, and therefore of *B*. A comparison of this difference (*A*'s point of view) with the difference between

B's own appearance and recognition patterns (*B*'s view of the relation) will trigger a particular emotional state in *B* that will have an impact on *B*'s physiology, and produce an appropriate behavioural reaction towards *A* intended to negotiate the difference between their images. The recognition process does not need to be cognitive or rational itself; indeed, other phenomena such as empathy or what has been called sympathetic coupling (Numaoka 1997) play an important role in emotion recognition and emotion-based communication. For instance, a mechanism similar to the use of pheromones for rank and emotion recognition in some mammals can serve to implement the recognition of appearance and presentation, and the effects that their difference produces in the agents' physiology and behaviour. In cases when it is important to maintain a history of individual interactions with other agents (or of interactions with types of agents belonging to different cultural categories or fulfilling diverse social roles), this difference in viewpoints can be kept as an emotional memory or snapshot that bootstraps the process in future interactions.

6. Conclusion and Prospect

The main roles that emotions play in social interactions in COMRIS can be summarised according to different scenarios: a synchronous and an asynchronous one. At the individual agent level, emotions ground the notions of competence and relevance. Emotional states being in part a consequence of the agent's competence and relevance, as reflected in the level of the pertinent physiological parameters, these two notions are manifested in terms of emotional expression. This way, they have a direct impact in social interaction. In turn, the level of success in this interaction feeds back into the agent's emotional state.

In the synchronous scenario, emotional states play a major role in the agents competition for the user's attention. Similarly, emotional expression is used by the parrot to solve the competition for attention game. Ideally (but admittedly not easy) this should imply the selection of agents that seem to best fit not only the current interests of the user (e.g., the agents that look more self-confident) but also her emotional state (e.g., the parrot may decide that a sad user might need a happy agent).

In addition to the above-mentioned mechanism of competition for atten-

tion, we are also starting to focus on the longer term, asynchronous coordination of agents' behaviours. If, as we argue, social interactions are based on constellations of agents' interests that can be furthered by those interactions, then long-term social structures must rely on some explicit or implicit encoding of these interest relationships. Instead of a purely cognitive encoding in which the agent accumulates a memory of other agents' (perceived or communicated) interests, we propose to explore alternatives that are no doubt being used in real societies.

In the asynchronous scenario, we envisage two main roles for emotions. One is the negotiation of difference between individuals — of their proximity in terms of interests and of the symmetry of their relation as perceived by each partner. An interesting issue to explore here is whether, and under what circumstances, the local interactions between couples (or small groups) of agents can lead to the equilibrium of the society, as claimed by Aristotle for his City, i.e., the political role of emotions. The other role is in the creation of long-term interest relationships. For example, a very self-confident agent unable to achieve a goal on its own may have to call for another agent's help. This can make the first agent to feel ashamed because it was not self-sufficient, and create a debt towards the second agent that it might want to pay back only privately — without the presence of witnesses that would know about its weakness. Or a sad agent that received the gracious help of another one can develop a long-lasting gratitude debt towards it, possibly implying more than one action at different times to show (and pay back) this gratitude. A concrete case that we have explored in more detail — and which is also well studied in anthropological literature — is the ritual act of present giving (Keane 1997). A present, whatever physical object it is, becomes a re-presentation of the present-giving act, which is embedded in a complex socially determined script of recognition and obligation. By its persistent nature and the emotional binding that both giver and receiver have with it, it creates co-behavior patterns that can extend over longer periods of time, especially when the exchange was witnessed by others (think of marriage rituals). We are proposing a scheme in which agents can exchange material tokens, that they associate with an emotional snapshot of the exchange event. In later dealings with the other agents involved, the token-snapshot combination triggers a (culturally determined) emotional pattern.

Another aspect of our work, more applied, is the development of technology for realising the various forms of coupling that we have proposed in a

society when viewed as a cohabited mixed reality. Surely, high-end information technology can provide interesting tools for this. The COMRIS parrot is one example that we hope to demonstrate in the next few months. We are also working on extending the awareness capability to include the activities and physiological state of the user. The COMRIS parrot, at that point, will become an example of an affective wearable (Picard 1997), that can take into account the affective state of its host, to better interact with other real or synthetic agents, along the lines explained in this paper. In this way, then, we hope to give bodily-grounded emotions their social role, not only in human societies but also in virtual and hybrid ones.

Acknowledgements

We are indebted to Kerstin Dautenhahn and to many participants in the 1997 AAAI Fall Symposium *Socially Intelligent Agents* for their valuable comments and feedback on an earlier version of this paper. While working on this paper, Dolores Cañamero was a postdoctoral fellow at the VUB AI-Lab. Support for her research was in part provided by the Spanish Ministry of Education (MEC) under grant PF95–00410164, and in part by the VUB AI-Lab. This work is currently supported by the EU-sponsored COMRIS project (LTR 25500) in which both IIIA-CSIC and Riverland Next Generation are partners. Other partners are the Vrije Universiteit Brussel (B), University of Reading (UK), Institute for Perception Research (NL), GMD (D) and the University of Dortmund (D). Walter Van de Velde is a honorary senior researcher for the Flemish Science Foundation (FWO).

Notes

1. An earlier version of this paper is available as Cañamero, D. and Van de Velde, W. 1997. "Socially Emotional: Using Emotions to Ground Social Interaction." In *Socially Intelligent Agents*, AAAI Press Technical Report FS—97—02, 10—15, © 1997 American Association for Artificial Intelligence.

2. COMRIS (LTR 25500, 97–2000) is a project funded by the European Union within their Long-Term Research initiative Intelligent Information Interfaces (I3). COMRIS stands for Co-Habited Mixed Reality Information Spaces. Its WWW home page is at http://arti.vub.ac.be/~comris. The views in this paper are not necessarily those of the COMRIS consortium.

References

Armon-Jones, C. 1986. The Social Functions of Emotions. In R. Harr (Ed.), *The Social Construction of Emotions*. Oxford: Basil Blackwell, 57–82.

Ashby, W.R. 1952. *Design for a Brain*. London: Chapman and Hall.

Averill, J.R. 1980. A Constructivist View of Emotion. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, Research and Experience, Vol. 1*. New York, NY: Academic Press, 305–339.

Cañamero, D. 1997a. Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In W. Lewis Johnson (Ed.), *Proceedings of the First International Conference on Autonomous Agents*. New York, NY: ACM Press, 148–155.

Cañamero, D. 1997b. *A Hormonal Model of Emotions for Behavior Control*. VUB AI-Lab Memo 97–06, Vrije Universiteit Brussel, Belgium.

Cornelius, R.H. 1996. *The Science of Emotion*. Upper Saddle River, NJ: Simon & Schuster.

Damasio, A.R. 1994. *Descartes' Error. Emotion, Reason, and the Human Brain*. New York, NY: G.P. Putnam's Sons.

Dantzer, R. 1994. (2nd ed). *Les Emotions*. Paris: Presses Universitaires de France.

Dautenhahn, K. 1996. Embodied Cognition in Animals and Artifacts. In M. Mataric (ed), *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Synposium*. AAAI Technical Report FS–96–02. Menlo Park, CA: The AAAI Press, 27–32.

Frijda, N.H. 1995. Emotions in Robots. In H.L. Roitblat and J.-A. Meyer (eds), *Comparative Approaches to Cognitive Science*. Cambridge, MA: The MIT Press, 501–516.

Genesereth, M. & S. Ketchpel. 1994. Software Agents. *Communications of the ACM* 7(37): 48–53.

ISWC'97, *Proceedings of The First International Symposium on Wearable Computers*. Los Alamitos, CA: The IEEE Computer Society Press, 1997.

Kalin, N.H. 1997. The Neurobiology of Fear. *Scientific American*, Special Issue 7 (1): 76–83.

Kandel, E.R., J. H. Schwartz, T. M. Jessell. 1995. *Essentials of Neural Science and Behavior*. Norwalk, CT: Appleton & Lange.

Keane, W. 1997. *Signs of Recognition. Powers and Hazards of Representation in an Indonesian Society*. Berkeley, CA: University of California Press.

Kemper, T.D. 1978. *A Social Interactional Theory of Emotions*. New York: Wiley.

Kemper, T.D. & R. Collins. 1990. Dimensions of Microinteraction. *American Journal of Sociology* 96(1): 32–68.

Kemper, T.D. 1993. Sociological Models in the Explanation of Emotions. In M. Lewis and J.M. Haviland (Eds.), *Handbook of Emotions*. New York: The Guilford Press, 41–51.

LeDoux, J. 1996. *The Emotional Brain*. New York, NY: Simon & Schuster.

Lutz, C.A. 1988. *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and their Challenge to Western Theory*. Chicago: University of Chicago Press.

Maturana, H. & F. Varela. 1987. *The Tree of Knowledge*. Boston, MA: New Science Library.

Newell, A & H. Simon. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Numaoka, C. 1997. Innate Sociability: Sympathetic Coupling. In K. Dautenhahn (ed), *Socially Intelligent Agents: Papers from the 1997 AAAI Fall Symposium*. AAAI Technical Report FS–97–02. Menlo Park, CA: The AAAI Press, 98–102.

Picard, R.W. 1997. *Affective Computing*. Cambridge, MA: The MIT Press.

Pribram, K, H. 1984. Emotion: A Neurobehavioral Analysis. In Scherer, K.R. and Ekman, P. (Eds.), *Approaches to Emotion*. Hillsdale, NJ: Lawrence Erlbaum, 13–38.

Sawhney, N. & C. Schmandt. 1998. Speaking and Listening on the Run: Design for Wearable Audio Computing. In *Proceedings of the Second International Symposium on Wearable Computers (ISWC'98)*, Los Alamitos, CA: IEEE Computer Society Press.

Tomkins. S.S. 1984. Affect Theory. In Scherer, K.R. and Ekman, P. (Eds.), *Approaches to Emotion*. Hillsdale, NJ: Lawrence Erlbaum Associates, 163–195.

Van de Velde, W. 1997. Co-Habited Mixed Realities. In F. Hattori (Ed.), *Proceedings of the IJCAI'97 Workshop on Social Interaction and Communityware*. Nagoya, Japan, August 25, 1997.

Van de Velde, W., S. Geldof, & R. Schrooten. 1998. Competition for Attention. In M.P. Singh, A. Rao, M.J. Wooldrige (Eds.), *Intelligent Agents IV: Agent Theories, Architectures, and Languages (Proceedings of the ATAL'97 Workshop)*. LNAI Series No. 1365. Berlin and Heidelberg: Springer-Verlag, 297–311.

CHAPTER SEVEN

Architectural Requirements for Human-Like Agents Both Natural and Artificial

What sorts of machines can love?

Aaron Sloman

1. Can Machines have Emotions?

In February 1998 I was invited to a literary society to talk on whether machines can love. The presentation was a mixture of philosophy of mind, literary quotations on love, speculation about evolution, theoretical ideas from Artificial Intelligence, and conjectures about human minds. Later Kerstin Dautenhahn kindly invited me to convert my slides into a chapter for this book. The result is a collection of conjectures about information processing mechanisms underlying human emotions, moods, attitudes and other cognitive and affective states, like love and grief. I shall provide some sketchy evidence that both common sense and the work of poets and playwrights involve an implicit commitment to an information processing infrastructure. However, other things besides healthy adult human beings have minds, and different sorts of minds require different sorts of information processing architectures.

If we analyse familiar mental states and processes found in normal adult humans, and compare them with capabilities of infants, people with brain damage or disease, and other animals, we find evidence for a diverse array of

architectures each supporting and explaining a specific combination of mental capabilities. This provides a broader and deeper explanatory theory than is normally found in philosophy or psychology. It also requires going beyond the majority of AI projects in considering both designs for *complete* agents and also *comparative* analysis of different sorts of designs as suggested in (Beaudoin & Sloman 1993; Sloman 1993; Mithen 1996).

No amount of observation of the behaviour of any animal or machine can determine the underlying architecture, since in principle any lifelong set of behaviours can be produced by infinitely many different information processing architectures. We can attempt to constrain our theories by combining a number of considerations, such as: (a) trade-offs that can influence evolutionary developments, (b) what is known about our evolutionary history, (c) what is known about human and animal brains and the effects of brain damage, (d) what we have learnt in AI about the scope and limitations of various information processing architectures. I offer a brief and incomplete report on a theory based on such constraints. The main conjecture is that human information processing makes use of (at least) three different concurrently active architectural layers which evolved at different times, which we share with other animals to varying degrees, and which, along with various additional supporting modules, account for different cognitive and affective states, as well as offering the hope of explaining different kinds of learning and development, different possible effects of brain damage, and other abnormalities. Such an architecture could give robots human-like mental states and processes.

Prejudice about machines and information processing often gets in the way of understanding and evaluating such theories, so that people ignore some rich explanatory ideas developed in the last few decades (e.g. Herbert Simon's important ideas (Simon 1967)). I shall therefore sketch and comment on the two main kinds of resistance to these ideas: doubting and fearing.

2. Doubters and Fearers

Many people are sceptical about or disturbed by the idea that robots or software agents may one day have thoughts, feelings, hopes, ambitions and the like, or experience the world as we do. Some are influenced only by evidence, others by fear, or dislike.

2.1 *Doubters: the perceived gap*

Many are doubters because they see the limitations of existing computer-based machines and software systems and cannot imagine any ways of overcoming these limitations. They do not realise that we are still in the early stages of learning how to design information processing systems.

Existing AI systems do not yet have whatever it takes to enjoy or dislike doing something. They do not really *want* to do something or *care* about whether it succeeds or fails, even though they may be programmed to give the superficial appearance of wanting and caring. The attempts to replicate other animal abilities are also limited: for example, visual and motor capabilities of current artificial systems are nowhere near those of a squirrel or nest-building bird, as I have argued in (Sloman 1989).

Because of the huge gap between machines developed so far and what animals can do, some people think the gap can never be bridged. That could turn out to be correct, if, for instance, the functioning of animal brains turned out to require some kind of mechanism that we have not yet dreamed of. The question is open.

It may be possible to convince some doubters by (a) enhancing their understanding of the real but unobvious possibilities of information processing machines, and (b) deepening their understanding of our ordinary concepts of 'feeling', 'thought', 'desire', 'love', etc., in order to reveal how our ordinary concepts of mind implicitly presuppose an information processing substratum.

Often defenders of AI do only (a). They try to remove doubts by demonstrating sophisticated things computers can already do, and pointing out that their capabilities will be enhanced by faster processors and bigger memories. That often fails to convince because it does not address the nature of mentality. Only by providing new insights into mental phenomena can we hope to convince real doubters that processes in computers may one day include feelings, experiences and thoughts. I shall sketch an attempt to bridge that gap below.

2.2 *Fearers: The longed for gap*

Some who reject the idea that robots and virtual agents can think and feel simply do not *like* the idea of machines (as they construe them) ever being so much like us. They may dislike it for many reasons, including fear of ma-

chines taking control (as in many science fiction novels) or more subtly because like Weizenbaum (1976) they fear that somehow human dignity is threatened if 'mere machines' turn out to be capable of all the interesting and important mental processes for which we value humans.

This kind of ontological neurosis (excessive concern about the place of humans in the overall scheme of things) lay behind at least some of the opposition in the past to the Copernican theory, which pushed us from the centre of the universe, and to the Darwinian theory of evolution, which blurred cherished boundaries between humans and other animals, a continuing concern of many researchers into animal capabilities.

In this paper I ignore the fearers who *dislike* the idea that robots will one day turn out to be like us. Dealing with such worries requires more than argument. Pointing out that intelligent machines could hardly do more horrible things to humans than humans do to one another is unlikely to help. I shall also not discuss theological objections, since I think they are based on false premisses.

2.3 *Ask how, not whether*

Whether machines can think, feel, care, hope, learn, have emotions, etc. is not in question, for humans are machines, though not artefacts. What sorts of machines?

3. Four Kinds of Machines

There are at least four kinds of machines known to science and engineering. They are not mutually exclusive: the same thing can be in two or more categories.

a. *Machines which manipulate force and energy.*

These include many machines that people (and some animals) have made for centuries, including many kinds of tools.

b. *Machines which manipulate matter by reorganising it.*

These include diggers, lawn-mowers, nut-crackers, looms, moulds, and also chemical and biological mechanisms which decompose and reorganise matter at the atomic or molecular level, for instance in production of solvents, detergents, drugs, etc. Every biological organism both transforms forces and

energy and also uses matter-transforming machines which take in nutrients and manufacture tissues, hormones, blood cells, sperm, and so on. Many physical machines are simultaneously of types (a) and (b).

c. *Machines which transform physical state.*

These include ovens, forges, and many machines in chemical plants. At the molecular level they can be viewed as a special case of (b).

d. *Information manipulating machines.*

These acquire, create, store, transform, manipulate, use and transmit information. Exactly what this means is a very subtle and complicated topic, discussed in (Sloman 1996a; Sloman 1996b). Information manipulating capabilities cannot exist without being implemented in a physical machine. In philosophers' jargon, information processing capabilities are *supervenient* on physical capabilities.

Organisms are not simply machines which manipulate forces and energy, and transform matter: they are also information processing machines. However, there is much diversity. They obtain information from the environment in different ways, store it, use it, transform it and communicate in different ways. They also deal with different kinds of information. An earthworm has neither the need nor the ability to know where Paris is, or how to multiply two numbers.

Humans depend on a mixture of mechanisms dealing with different sorts of information, processed in diverse ways, including sensing the environment, learning a language, absorbing a culture, generating new goals, making plans, evaluating and selecting plans, learning skills, learning generalisations, and many more. As Wiener noted, many of these processes are primarily concerned with *control*, e.g. control of attention.

Many people think of 'information processing' as restricted to computers manipulating bit-patterns in rigidly programmed ways, e.g. (Rose 1993). This can lead to spurious arguments against information processing models of minds, or brains. We require a broader notion of 'information processing,' as used by many software engineers, biologists, and some brain scientists (e.g. Damasio 1994).

3.1 *Poets on love*

I shall try to show how being an information processor is involved in many mental states, e.g. loving and fearing. For instance, Shakespeare was implic-

itly alluding to features of an information processing system when he wrote:

> LOVE IS NOT LOVE
> WHICH ALTERS WHEN IT ALTERATION FINDS

This implies that lovers can find alteration, i.e. perceive changes in loved ones. Finding alteration often diminishes or wipes out love and trust. Yet a common theme in literature is that true love is not so easily changed. It is resistant to common forms of information processing, including discovering disappointing facts about the beloved. Thus in love, some control states are unusually resistant to being changed by new information. There are many relevant entries on love in the *Oxford Dictionary of Quotations*, including humorous poetry which alludes implicitly to information processing mechanisms, for instance when Sir John Suckling jokes about the oft claimed constancy of love:

> OUT UPON IT.I HAVE LOVED,
> THREE WHOLE DAYS TOGETHER
> AND AM LIKE TO LOVE THREE MORE,
> IF IT PROVE FAIR WEATHER

Of course, I am not claiming that such authors had clear ideas about information processing, though aspects of the chemical infrastructure of our information processing are often acknowledged, e.g. when Calverly wrote:

> THE HEART WHICH GRIEF HATH CANKERED
> HATH ONE UNFAILING REMEDY — THE TANKARD

I shall try to show that some of the information processing capabilities of most interest to us in our social life (including the ability to be in love) depend on aspects of our architecture which evolved recently and are probably not shared with most animals, except perhaps other primates (though I am not sure). We also have much older information-based control mechanisms which are shared with many other animals and which are easier for brain scientists to study (LeDoux 1996). These explain some of our more primitive emotions, as explained below.

I shall not attempt to *prove* that we are information processing machines, but will merely try to explain in what sense we are, with illustrations of information processing capabilities. The ultimate test of the idea will be our ability to develop more detailed theories which are better able to explain the

full range of human and animal mental capabilities than rival theories. That will take some time!

3.2 *Cluster concepts cannot be defined precisely*

Some readers may hope for definitions of terms like *information processing*, *mental process*, *consciousness*, *emotion*, *love*. However, each of these denotes a large and ill-defined collection of capabilities or features. There is no definite collection of necessary or sufficient conditions (nor any disjunction of conjunctions) that can be used to define such terms. The features and capabilities involved in mentality or consciousness or emotions can be present or absent in different combinations, in different animals, in people at different stages of development or after brain damage. (And, some of us claim, also in future robots.) Such concepts can be described as 'cluster concepts.' (Compare family resemblance concepts discussed in (Wittgenstein 1953)).

If *emotion* is a cluster concept it is a mistake to ask how *it* evolved, what *its* function is, what *its* neural correlates are, etc., for there is no definite *it* to which the questions relate. (The same applies to *consciousness*.) I shall illustrate this below in relation to emotions by showing how different architectures support different subsets of the loosely defined cluster of features associated with our ordinary notion of *emotion*.

However, if the phenomena are all related to some general underlying principles, such as principles common to different information processing architectures, then it may be possible one day to define precise new technical concepts in terms of those principles. E.g. below I shall (partially) define 'tertiary emotion' in terms of a type of architecture. Such new theory-based concepts are often loosely related to pre-scientific cluster concepts which inspired the new theories. This has happened many times in the history of science, including refinement of our pre-scientific concepts of kinds of stuff, of kinds of animals, and kinds of chemical processes. It is only *after* we have deep explanatory theories that precise definitions can be given. Unfortunately we unwittingly deceive ourselves into thinking that we start with clear and precise concepts, e.g. of *experience*, *emotion*, etc. Likewise people thought they had a clear and precise concept of simultaneity, until Einstein exposed the problems.

4. The Sorts of Machines We Are

Until recently the only significant information processing machines were organisms. However, since the middle of the 20th century our understanding of and ability to create new kinds of artificial information processing machines has accelerated rapidly, though the science is still in its infancy, and we have much to learn. In this and the next two sections I shall elaborate, first in very general terms, then in more detail, an information-processing model of mind. I assume that we are physical, chemical, biological and information processing machines:

- rooted in carbon, hydrogen, oxygen, nitrogen, iron and other physical stuff,
- evolved through millions of years of exploration,
- partly revealing our history in our design,
- grown in wombs, cots, playgrounds, and cultures,
- acquiring, storing, transforming and using energy,
- acquiring, storing, transforming and discarding matter,
- acquiring, storing, transforming and communicating, information,
- using information in many ways, including sensing, deciding, doing and feeling,
- writing poems, plays and newspaper reports,
- providing the stuff to write about,
- deceiving ourselves that we are unique,
- often wanting the truth to be THUS ..., rather than wanting to know what the truth is!

4.1 *But we are not 'just' machines*

We are machines but we are not *just* or *mere* machines, any more than computing systems are. Computing systems are certainly information processing machines, but they can also be personal assistants, factory controllers, tutors, translators, planners, network managers, automatic pilots, theorem provers. Beware the *nothing buttery* fallacy: the temptation to conclude that something is 'nothing but a ....'.

A system may be describable using a certain ontology without that being all there is to the system, even if the description is *complete* at that level. The people, buildings, transport mechanisms, etc. in a large city can (in principle) be described in very great detail using the language of physics and chemistry.

But that does not mean there is nothing else, for there will probably also be crime, poverty, jobs, salaries and laws, obligations, contracts, and knowledge. Moreover these non-physical entities can have important causal powers, despite the causal completeness of the physical level.

In philosopher's jargon (explained at length in Chalmers 1996), many non-physical entities, like crime and poverty, are 'supervenient' on the physical infrastructure. But that does not stop them being real and causally efficacious. Poverty *really* can cause crime and in doing so it can cause physical events, like TV sets moving through broken windows and knives or bullets through skin. Likewise events in an information processing *virtual machine* can cause events in a physical machine: for instance changes on a computer screen, movements of a robot, the safe landing of an airliner, and temperature control in a hospital intensive care ward.

If a physical description of a system is complete at its own level people sometimes infer that the system is 'nothing but' a collection of atoms, molecules, etc.. This is the 'nothing buttery' fallacy. An ocean wave might seem to be nothing but a large collection of molecules partaking of roughly vertical or circular motion at a fixed location, but that ignores the large scale horizontal motion and forces which can have such destructive effects when they hit the shore. Interactions between levels in information processors are more subtle: physical events cause virtual machine events and some virtual machine events cause physical events.

Some multi-level systems are much harder to understand because they are so subtle and complex, and because we have not yet learnt the concepts and techniques required for thinking about how they work. Trying to describe precisely the relations between minds and brains may be premature at present, if we do not yet have sufficiently rich and subtle concepts. It is much easier to start from systems we have designed. After developing conceptual tools for explaining how they work we may be able to extend those tools to deal with more complex cases.

We understand how a typical computing system contains several levels of virtual machines within which causal interactions occur between information structures. Processes such as reformatting a document, or finding a logical proof, or interpreting an image, are real and efficacious, even though at a lower level the computer is completely describable in terms of its digital electronics. At a still lower level quantum physicists use yet another set of concepts. Maybe physicists of the future will find something even deeper. But

that need not affect the reality of the levels we now know.

To take any particular level and say: there is really nothing but *that* is to impose arbitrary constraints on what is real. We cannot ignore the existence and causal powers of poverty and crime. (Though some politicians may find it tempting.) Likewise a partially completed proof in a machine is real, and can cause both internal processing events and external physical events (on a screen) even though no physicist or electronic engineer could observe or measure the proof.

Lack of understanding of information processing virtual machines and how they can be implemented in physical systems has led many philosophers and theologians to assume that thoughts and feelings must inhere in some non-physical kind of mechanism, sometimes called a soul, or spirit, which can exist without any physical implementation. Gilbert Ryle scornfully labelled this 'The ghost in the machine'. His 1949 book had important ideas about internal, unobservable, information processing, e.g. in his chapter on imagination. But he lacked our conceptual tools, and as a result was wrongly interpreted as a behaviourist, denying the existence of the mental.

4.2 *We are multi-level information processors*

Although we have physical architectures we are not 'mere' physical machines. We also have information processing architectures, implemented in our physical architectures. Likewise we are not 'mere' information processors since we are also parents, teachers, criminals, lovers, scientists, etc. These features and relationships are implemented in our information processing capabilities in combination with a social context.

Information processing includes: sensing the environment, interpreting sensory information, modifying stored information (beliefs, desires, intentions, plans, skills) in the light of new information, generating goals by various means (some innate, some learnt, some unconscious, some conscious), inventing new options (things to do, to make, to look for...), considering and evaluating options, selecting among possible actions, wondering about consequences, reconsidering previous decisions, and much more. Analogous capabilities exist in sophisticated game playing machines.

Humans also enjoy some activities and dislike others, (sometimes) detect our own states (anger, puzzlement, hope, ...), evaluate our own thoughts and reasons (as selfish, unproductive, altruistic, creative, foolish, etc.), feel

ashamed, guilty, fearful, excited, and become self-satisfied, infatuated, obsessed, ecstatic, ... Will AI systems ever have all these capabilities? I see no reason to doubt the possibility (or even to fear it). Some counter-arguments seem to be born of dislike of the idea (a 'longed for gap') rather than deep analysis of what it is to enjoy something or be infatuated.

Not all animals can do all those things. Not all humans can do them all: very young children, and people whose brains are either genetically flawed or damaged by accident or disease, may lack some of these abilities. We need to understand why.

4.3 *Longing as a tertiary emotion*

Consider an example: *Why can't a goldfish long for its mother?* Longing for one's mother involves at least: (i) knowing one has a mother, (ii) knowing she is not present, (iii) understanding the possibility of being with her, and (iv) finding her absence unpleasant. These all involve possessing and manipulating information, e.g. about motherhood, about one's own mother, about locations and change of location, and about the desirability of being close to one's mother.

Those conditions do not suffice. If someone in Timbuctu whose mother is in Montreal satisfies conditions (i) to (iv) but hardly ever thinks about his mother, and simply gets on with his job, enjoys his social life, and always sleeps soundly, then that is not a case of longing. He may *regret* her absence (an attitude), but he does not *long* for her (an emotion). Longing for someone requires something more, namely (v) not easily being able to put thoughts of that someone out of one's mind. Thoughts of the longed for one will return willy nilly. (Though perhaps not in *mild* longing!) This is not just a matter of definition: it is a fact that some human mental states involve partial loss of control of attention.

You cannot lose what you have never had. So a requirement for being in such a states is having the *ability* sometimes to control what one is thinking of and also being able sometimes to *lose* that control. This presupposes an information processing mechanism some part of which can control which information is being processed, but which is not always in *total* control. Deep longing for one's mother involves partly losing control of thought processes, a *perturbant* state. An incomplete grasp of these ideas gives rise to a confused notion of *free will* which appears to some people to be inconsistent with

causation (contrast Franklin 1995).

I call these perturbant states 'tertiary emotions', as explained below. Other examples are anger (Sloman 1982), grief (Wright, Sloman & Beaudoin 1996), guilt, jealousy, excited anticipation, infatuation and many others. We shall see below that such perturbant states arise in an architecture where a high level control mechanism sometimes loses control. That capability is not one of its functions but is a side effect of other functions. Such states could also occur in some intelligent robots (Sloman & Croucher 1981).

4.4 *What sort of architecture could support being in love?*

There are many other states which characteristically involve partial loss of control of attention. This is one of the differences between loving someone and being in love.

X is in love with Y IMPLIES *X's thoughts are constantly drawn to Y*

Love in general is an *attitude* and need not be emotional: You can love members of your family without constantly dwelling on them. You can also love your country, love the organisation you work for, love football, love the music of Mozart, without any of these constantly flooding your thoughts.

Loving your country does not involve thinking about it most of the time, but only when some relevant information or decision turns up. The rest of the time the love is just one among many *dormant dispositions* — but real all the same. Being *in love* is not so passive: thoughts of the beloved will return when there is no particular reason. In extreme cases (infatuation) it may be very difficult to think about anything else. Likewise in extreme grief. However, even extreme emotions can temporarily become dormant while some urgent and important task, or a gripping movie, holds one's attention. To explain all this we need to understand the underlying information processing architecture which makes attending and thinking possible at all, and which accounts for the possibility of redirection of attention. Explaining how we can *lose* control of our thoughts first involves explaining how we can *have* control.

5. Architectural Layers

5.1 *An architecture explains a collection of states and processes*

A particular information processing architecture will support some states and processes, but not others. Some problem-solving processes require an architecture including a procedure invocation stack, providing an ordered 'memory' of unfinished procedures. A condition-action rule interpreter with no explicit stack makes it hard to implement strategies requiring deeply nested actions, although it provides good support for opportunistic information processing. An architecture which includes both a stack and mechanisms for explicitly inspecting or changing its contents, will be less restrictive than one without.

An architecture in which there is always only one process active makes it hard to implement self-monitoring and self-control, whereas an architecture supporting (physical or virtual) concurrent processes makes it easier for one process to inspect the state of another, and interrupt or modulate it, e.g. if looping is detected.

An information processing architecture explains a variety of states and processes somewhat as the atomic theory, a theory of the architecture of matter, generates and explains a variety of types of physical elements and chemical compounds.

Our knowledge of information processing architectures is still very primitive. Studying a wider range of architectures will extend our ability to explain how different collections of competences are possible. Each architecture provides a framework for generating a family of descriptive and explanatory concepts. We can expect to find different architectures in different sorts of humans (including infants and people with brain damage, etc.), different sorts of animals and different sorts of artificial agents.

5.2 *A conjectured architecture for adult humans*

Within an architecture we can distinguish perceptual subsystems, motor subsystems and more central mechanisms. I conjecture that in adult humans all of these consist of layers with different levels of sophistication which evolved at different times, and which are shared with different numbers of other animal species, and explain different aspects of human mentality, for instance different types of emotions. The layers act in parallel and both cooperate and

compete with each other. The different 'layered' capabilities in sensory and motor subsystems evolved to work with the different central layers. All this has some similarities to many other theories, e.g. (Craik 1943; Minsky 1987; Damasio 1994; Dennett 1996; Mithen 1996).

5.3 *Three types of sub-architecture*

The central layers are (1) a very old *reactive* layer, found in all animals, including insects), (2) a more recently evolved *deliberative* layer, found in varying forms in a subset of other animals, (3) an even more recently evolved *meta-management* layer providing self-monitoring and self-control, perhaps found only in other primates, and probably not in very young human infants. I.e. the architecture of an adult human is not present at birth, but results from a boot-strapping process. Each layer is a collection of cooperative sub-mechanisms combining to perform a collection of internal or external functions. Additional modules support or modulate the three main layers:

a. One or more global *alarm* systems able to detect patterns requiring rapid global reorganisation of internal and external behaviour. Compare the interrupt mechanisms discussed by Simon (1967) Oatley and Johnson-Laird (1987), and the role of the amygdala in the theories of LeDoux (1996), Damasio (1994).

b. Associative content-addressable information stores, essential for 'what if...' deliberations, and for predicting what is likely to happen next (Craik 1943).

c. Mechanisms generating, comparing, selecting and prioritising motives (Simon 1967; Sloman & Croucher 1981; Beaudoin 1994; Beaudoin & Sloman 1993).

d. Global quantitative and qualitative control subsystems which account for mood changes, differences between waking and sleep, types of arousal, etc. Some of these global controls in brains use chemical mechanisms, but similar functions might be implemented differently in artificial agents.

5.4 *Layered perceptual and motor systems*

Perceptual mechanisms need varying degrees of sophistication for their tasks. The contribution of vision to posture control uses relatively simple optical flow detection. Segmentation and recognition of objects in a scene requires more global and knowledge-based processing. Seeing a room full of people as

a 'party' or a 'seminar' or as 'highly charged' requires far more abstract and sophisticated forms of processing. Learning to read text or sight-read music involves different collections of layered abilities (Sloman 1989).

Some action mechanisms are old and relatively primitive, such as contracting muscles, raising or lowering a leg, clamping jaws shut. More abstract actions involve considerable sensori-motor coordination such as picking up a large, heavy, unwieldy and unfamiliar object. Some actions use tools as an extension of the body, such as parking a car or feeling the shape of a hole with a probe. There are also semantically very rich actions such as uttering a sentence, playing a musical phrase on the violin, or making courtly gestures. The 'layering' of sensory and action systems is obscured by thinking of such systems only as input and output transducers.

5.5 *Evolution by copying and modifying*

Some of the mechanisms might have evolved from simpler mechanisms via the common biological process of copying then modifying. For instance deliberative mechanisms used for planning require an associative store of information about actions possible in various situations and their consequences. This might have evolved by copying an older reactive association mechanism. The new mechanism instead of merely reacting to its input by producing *control* signals ('do this next') might have answered 'what if' questions about past actions, as part of a learning and debugging mechanism. Later it could answer 'what if' questions about future actions. (Or perhaps planning came first?)

All this requires an ability to synthesise hypothetical context descriptions to feed into the memory, instead of using only current sensory information to drive it. That might have resulted from earlier developments producing high level perceptual mechanisms able to create abbreviated abstract descriptions of external objects or situations.

Such developments could involve multiple evolutionary stages, not yet understood. Perhaps a self-monitoring meta-management mechanism was evolved by copying the global alarm system and then changing its activities, including making it more amenable to rule-based control, and allowing it to use deliberative strategies. This would be much slower and more flexible than an alarm system. Moreover, with rule-learning, it can improve itself, and be influenced by a culture.

5.6 *Different layers explain different sorts of emotional processes*

We can distinguish at least three major categories of emotions, explained by the three sorts of processing layers.

1. *Primary emotions* (Damasio 1994; Picard 1997). These are primitive emotional states (like being startled, terrified, sexually stimulated) based on the old reactive layer and global alarm system shared with many other animals. Patterns in sensory inputs are detected by the global alarm system which rapidly sends out a wide range of control signals, some causing physiological changes producing or preparing for action. Compare robots programmed to 'freeze' as soon as a human gets dangerously close.

2. *Secondary emotions*. These states (like being anxious, apprehensive, relieved, pleasantly surprised) are generated in the deliberative layer, in which plans can be created and executed, risks noticed in advance, progress assessed, success detected, etc. They depend on the 'what if' representational capabilities provided by a deliberative mechanism. Some emotions (like relief that an accident was avoided) require counterfactual information about the past (what *might have* happened). An alarm system detecting a feature or pattern in the contents of current thoughts or problems can be triggered to produce a rapid global reaction, or change of state (e.g. producing nervousness and thereby more attention to detail, raised interrupt thresholds, more cautious movements, etc.). Detection of success, or receding danger could trigger reversion to a more normal state, as in relief.

Damasio calls emotions triggered by such cognitive processes 'secondary emotions'. In chapter seven (especially page 137) Damasio (1994) suggests that secondary emotions *always* activate the same physiological mechanisms as primary emotions (Picard uses the phrase 'sentic modulation'). However 'always' is an over-generalisation. There are considerable individual differences regarding whether secondary emotions triggered by cognitive processes produce physiological changes. The more important feature of secondary emotions is the ability of something like the alarm mechanism to redirect *mental* processes, so that, for instance, dangers and opportunities are noticed and appropriate actions considered. Moreover emotional maturity sometimes involves suppressing normal physical reactions and dealing with emotion-generating phenomena entirely mentally. (Not everyone can do this.) Purely mental processes need not be *cold* and *unemotional*. On the contrary, they can be rich in evaluative content and powerful in their effects on other mental

processes, as when horrific news grips our attention despite our best efforts to think of something else.

We can therefore distinguish *purely central* secondary emotions from *peripheral* secondary emotions which invoke the bodily mechanisms used by primary emotions. Steve Allen alerted me to Damasio's chapter eight (page 184), where he explicitly adopts a similar position, suggesting that in some cases there are "as if" mechanisms which bypass the route through the body, so that "the prefrontal cortices and amygdala merely tell the somatosensory cortex to organise itself..." in a pattern that it would have assumed if signals had come through the body. (Damasio's ideas in that chapter have much in common with the ideas presented here, including his speculations about the role of ancient insect-like mechanisms in human brains.)

3. *Tertiary emotions*. These are typically human emotional states involving partial loss of control of thought processes (perturbance), e.g. states of feeling humiliated, infatuated, guilty, or full of excited anticipation, where attempts to focus attention on urgent or important tasks can be difficult or impossible, because attention is drawn back to the focus of the humiliation or infatuation, etc. This can happen despite a meta-management decision to attend to something else.

These fit the definition of secondary emotions, but involve something more, namely partial loss of control of attention. This is possible only if there is something which normally provides that control. Only then does the notion of 'losing control' become relevant. Without meta-management you cannot explicitly evaluate the possibility of attending to A and to B, and then decide to attend to B because you judge that better. Without meta-management, an alarm mechanism or other mechanism cannot undermine a decision based on an explicit judgement that it would be better to attend to or think about B. *Only if there is a control mechanism can control be lost.* Thus tertiary emotions require a particularly sophisticated information processing architecture.

Not *only* meta-management can redirect attention. Normal processes of deliberation involve shifting attention, e.g. switching attention to new goals, and switching from thinking about ends to thinking about means. Reactive mechanisms, e.g. detecting something bright or something moving, or detecting utterances of one's own name, can redirect perceptual resources. A deliberative system may have many sub-processes competing for attention, perhaps using a network combining connection weights and activation levels. A meta-management mechanism explicitly evaluates and selects alternative foci of

attention, recording that this has been done. Put anotherway: *just as external behaviour can be either reactive or deliberative, so can internal behaviour.* And in both cases the reactive mechanisms can sometimes defeat the deliberative mechanisms.

Further subdivisions are possible, of course. For instance, not all perturbances are emotional: emotions involve strong *evaluations* as well, and there are different sorts of evaluations: selfish, ethical, etc. I have no idea to what extent non-human animals (e.g. bonobos) have such self-monitoring and self-evaluating capabilities. They may have simpler versions supporting a different range of emotions involving some self awareness and self evaluation. This is a topic requiring empirical research and theoretical analysis of new architectures.

By using an architecture-based framework for defining different classes of emotions, and related notions like 'mood', 'desire', 'enjoyment' and 'pain', not discussed here, we avoid much argumentation at cross-purposes about emotions, because the theory supports a range of types. Dozens of different definitions of 'emotion' have been proposed. We can now see that there is no point arguing about which of N definitions is correct if there are N types of phenomena all of which need to be studied and explained. It does not matter whether we call them 'emotions' or not. The phenomena are important, not their names.

In various previous publications, e.g. (Sloman 1992), I focused mainly on tertiary emotions, because I thought they were of most interest and importance in human interactions. I mentioned other kinds e.g. reflexes and startles and also the types discussed by Oatley and Johnson-Laird (1987), but did not have a clear view of how the different types could fit into a common architecture.

6. Different Architectural Layers and Evolution

The three layers can now be described in a little more detail. Diagrams are used very impressionistically to indicate some of the features of the different mechanisms.

6.1 *Reactive mechanisms*

Reactive mechanisms evolved very early and are widespread in plants, insects and all other animals. The more recently evolved deliberative mechanisms

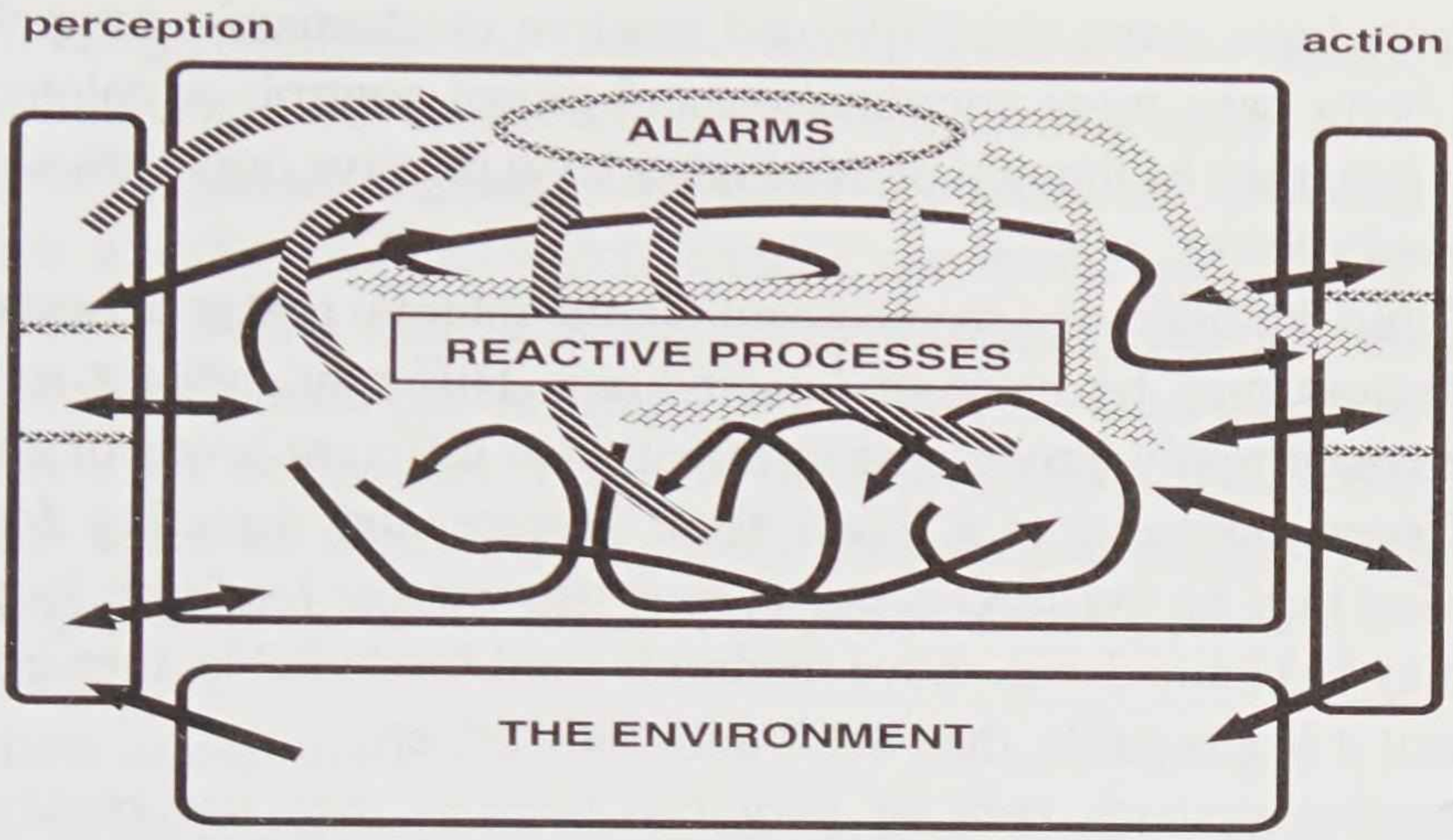


Figure 1. Reactive systems with alarms: emotional ants?
Internal reactive mechanisms may be chained together and may involve feedback loops, and changeable 'state variables'. Horizontal bars separate layers of abstraction in sensory and motor systems. Two-headed arrows indicate two-way flow, or feedback. Textured arrows represent fast links to and from the global alarm system.

have not *replaced* the older mechanisms, but function alongside them, but not necessarily always dominating them.

The main feature of reactive mechanisms is their *inability* to contemplate, evaluate, and select possible future courses of action or other hypothetical possibilities. They can merely react to actually detected internal or external situations. In other words they cannot consider novel options before selecting them, or create new plans. They merely act, though some of the actions are internal and some external. Figure 1 crudely depicts a fairly sophisticated reactive architecture, including alarm mechanisms, and the layered sensory and motor systems described below.

Reactive systems are very varied. They may be composed of purely analog (continuous), purely digital, or a mixture of analog and digital mechanisms. They may be totally *environment driven* so that the same environment always produces the same response, or partly *state driven* so that changeable internal states help to select actions. Detected needs can change internal states so as to modify subsequent selection among actions. (These are sometimes called *drives.*) This constitutes a primitive form of *goal-directed* though purely reactive behaviour. Compare Nilsson's Teleo-Reactive programs (1994).

Plants have many uncoordinated reactive mechanisms. Animals with a central brain have more coordination and global control. A colony of such animals can often be thought of as a higher level reactive organism with totally distributed control.

Within the brain of a reactive animal, the internal routes between sensors and effectors may be more or less indirect. Different indirect routes may operate concurrently, processing information at different levels of abstraction for different purposes (e.g. controlling posture and detecting food). The processing may be *uni-directional* or may use internal feedback loops. Some loops may be chained: e.g. move randomly until food visible, then go towards food until it is graspable, then eat it until satiated, etc.

Reactive systems may be very fast because they use highly parallel implementations. This may lead to simultaneous activation of different actions. Sometimes different actions can be performed in parallel, or combined, using something like vector addition (e.g. increasing speed and increasing angle of turn). If they are inconsistent a *selection* mechanism is required, e.g. symbolic priority rules, winner-takes-all neural nets, or a simple voting mechanism.

Layered sensory and action subsystems

Short horizontal lines in the diagram indicate divisions between sensory and motor processes operating at different levels of abstraction. For instance, some reactions depend on relatively simple measures of optical flow or contact pressure, while others use more sophisticated and global percepts produced by complex interpretive procedures classifying entities in the environment relevant to higher level decisions (e.g. recognising something as a shelter, or as dangerous). Likewise actions may be simple internal or external changes (e.g. contraction of a particular muscle) or more sophisticated hierarchically controlled actions. Feeding often requires high level coordination of limbs (grasping food) and jaws. Many forms of running, jumping, flying, nest-building and mating require very complex coordination of complex collections of muscles.

More flexible reactive systems

Some chained reactions involve innately determined sequences of internal states implementing plans selected by an evolutionary mechanism and encoded in genes. Others may be a result of new links produced by learning. Although purely reactive systems are rigid in that they cannot 'think ahead' creating and evaluating new plans, they may use learning mechanisms to alter

weights linking conditions and actions. Where some actions change internal states forming conditions for subsequent actions, learning can be used to chain sequences of responses to produce the effect of learnt plans, provided that the architecture already has links which learning can strengthen.

Further flexibility can be achieved by allocating internal storage for different contexts which can be turned on and off (or varied continuously), to modulate reactive behaviours. Yet more flexibility can be achieved by allowing internal reactions which create simple temporary structures, for instance representing goals (e.g. 'catch that animal', 'find a hiding place').

Global alarm *mechanisms*

If chains of internal reactions intervene between sensory input and corresponding output, this may sometimes cause fatal delays or missed opportunities. Some sort of global 'override' mechanism could deal with this: an 'alarm' mechanism which allows rapid redirection of the whole system in response to detected patterns indicating opportunities or dangers. The alarm mechanism, which might be either entirely innate or partly trainable, could be simply another reactive sub-system with inputs from all parts of the organism driving a fast trainable pattern-recogniser able to trigger outputs to all parts of the system. Normally it would do nothing, but when turned on by appropriate conditions it could rapidly redirect the rest of the organism to produce freezing, attacking, feeding, fleeing, mating, attending (sudden high alertness), more general arousal, or more specific innate and learnt responses.

It appears that such systems first evolved a long time ago: many animals have one or more global alarm mechanisms. The brain stem and the amygdala both seem to implement alarm systems which evolved at different times. Different global alarm mechanisms could specialise in particular types of activation patterns and response patterns. Primary emotions in vertebrates appear to be implemented in such systems (LeDoux 1996; Goleman 1996).

Robots and software agents do not yet have all the characteristics of the reactive architectures described here. However, there has been a lot of work in robotics labs on reactive systems (much of it inspired by Rodney Brooks at MIT), and it is very likely that more and more sophisticated insect-like, or lobster-like creatures will emerge from such laboratories in the next few years, and also software systems controlling chemical plants, power stations, etc., all with the capability to have the sorts of primitive emotions sketched here.

Whether they will *know* they have them, and whether ants fleeing 'in

terror' know they are terrified is another question. The third architectural layer sketched below can explain self-awareness. We may have to get used to the idea that without it reactions of terror and other primitive emotions may occur without being *experienced* as such by the organism. This could be equally true of new-born human infants, if they lack the third layer described below. The suggestion may seem repugnant, but that does not make it false.

6.2 *Architectures with deliberative mechanisms*

A deliberative mechanism provides capabilities missing from reactive mechanisms, especially the ability to achieve an objective, in a *new* situation, by chaining together a novel sequence of actions. A reactive system may be able to invoke an *existing* plan, e.g. if a need is detected and allowed to trigger a sequence of context-driven reactions. But that presupposes a pre-existing implicit or explicit plan, produced by evolution or previously learned chained responses.

Novel complex actions may be discovered by a reactive explorer using trial and error with reinforcement learning, but this can be dangerous and time consuming. If a system has the ability to do hypothetical reasoning it can search a space of *possible* action sequences until it finds a suitable plan, as Craik pointed out (Craik 1943). This requires a content-addressable associative memory store which can answer questions like: 'What actions are possible in situation X?' and 'What effects would follow if action A were performed in situation X?' 'Which actions are relevant to a goal of type G?'

A system able to create potential new plans to evaluate requires a re-usable memory in which to build partial plans before selecting them. The tree-like structures in Figure 2 indicate partially constructed possibly hierarchical solutions to problems. These are in the re-usable work space. This re-use will make the process of exploration serial. There are other reasons why deliberative mechanisms must be sequential (digital) and discrete, and relatively slow. For instance, even if the association store operates using a highly parallel and distributed neural implementation, it could still be restricted to answering one question at a time.

Extending the alarm mechanism

As before, alarm mechanisms may be useful for rapidly redirecting a deliberative system when dangers and opportunities are detected. States within the

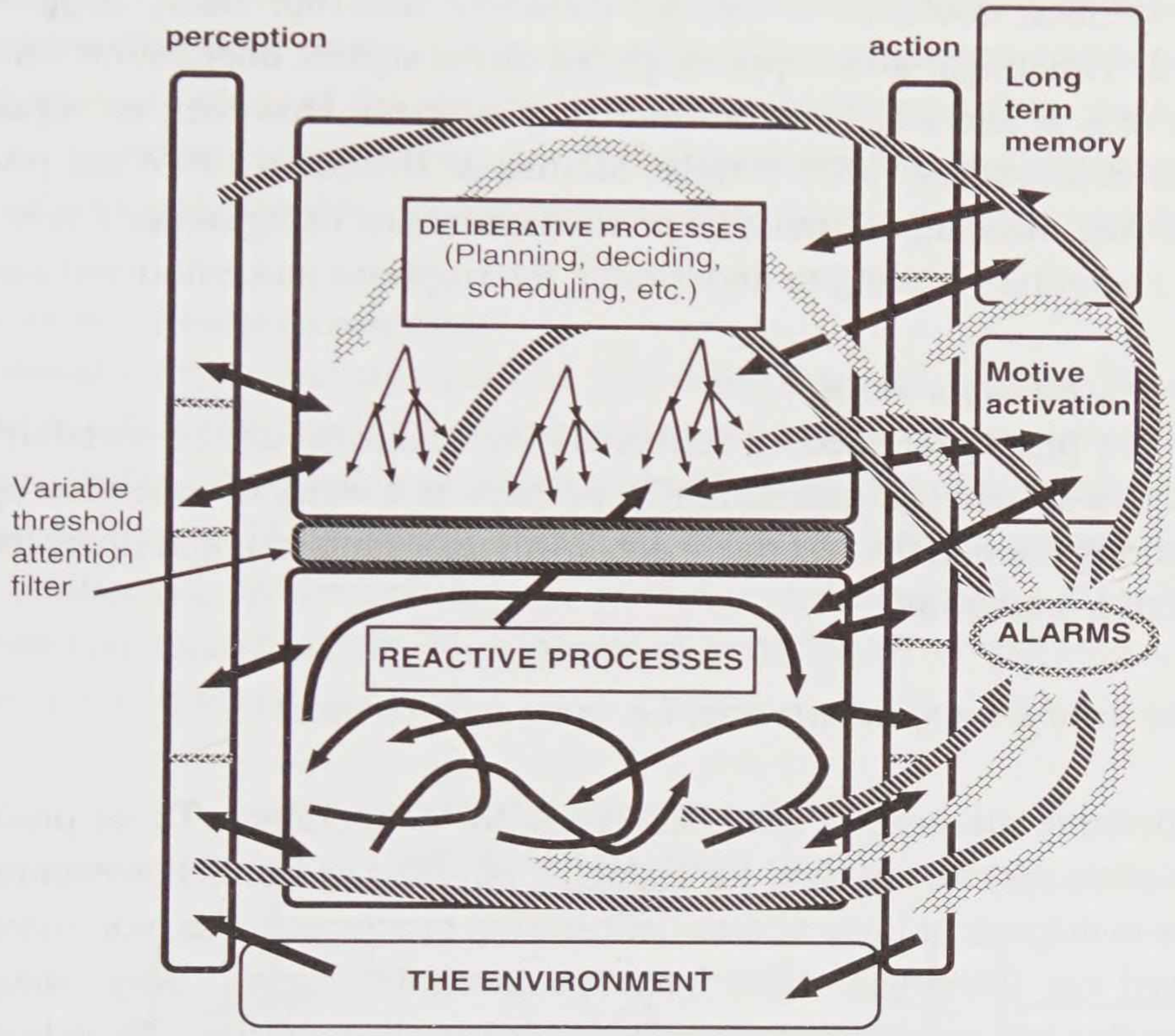


Figure 2. Hybrid reactive and deliberative architecture, with global alarms
Now the deliberative layer also has links to and from the alarm system. A filter with dynamically varying interrupt threshold protects the resource-limited deliberative layer when dealing with tasks that are important, urgent and resource consuming.

deliberative layer could also feed into the alarm system, alongside signals from sensors and reactive mechanisms. Similarly the alarm system could send interrupts and redirection signals to the deliberative mechanisms, re-directing attention or changing the mode of processing. This is indicated crudely in Figure 2.

An attention filter

A fast-changing environment can cause too many interrupts and frequent re-direction of attention, with more time spent switching between deliberative tasks than actually solving the problems (like a thrashing operating system). It may be important to prevent interruptions and diversions (e.g. by new goals) when the current goal is very important, urgent and cognitively demanding. A

partial solution could be a variable-threshold interrupt filter, depicted in Figure 2. This might also suppress global alarm signals under some circumstances (e.g. soldiers in battle not noticing injuries). However, as argued in (Sloman & Croucher 1981; Wright, Sloman & Beaudoin 1996) the priority and filtering mechanisms must be *fast* which means using unintelligent processes, sometimes leading to undesirable interruptions and emotional states.

Saving new plans for reuse

Useful new plans generated by deliberative mechanisms can be transferred to the reactive system (the cerebellum?), perhaps as a result of repetitive operation. Storing them in the reactive mechanism may support much faster though less flexible execution.

6.3 *The need for self-monitoring, i.e. meta-management*

A deliberative mechanism needs strategies for deliberating. Those produced by evolution may be too rigid for changing physical and social environments. A meta-management layer allows deliberation processes to be monitored and improved e.g. learning to raise interrupt thresholds during 'busy' states, or noticing that certain planning methods fail in certain conditions. Such learning may reduce failure in deliberative tasks, reduce interference between goals, detect time wasted on unsolvable problems, etc. Flexibility is even greater if meta-management can use rules, categories and values absorbed from the surrounding culture.

The ability to attend to and categorise internal states has subtle consequences, which may have influenced evolution of self-monitoring capabilities. Parents can diagnose a child's problems more easily if the child can attend to and describe internal symptoms. Compare describing visual experiences to an optician, or telling a dentist which tooth hurts. Attending to intermediate visual data-structures is required for drawing accurately: noticing how things *look* (e.g. elliptical) as opposed to seeing how they *are* (e.g. circular). (This could explain the existence of *qualia*.)

Further extension of the alarm mechanism

The alarm mechanism described previously could be extended with inputs from and outputs to meta-management processes, allowing alarm reactions to be triggered by and to modify meta-management. Alarm systems require rapid

reactions, so they must depend on fast, and therefore shallow, pattern recognition rather than deep analysis. Consequently, alarm processes will not always be optimal and some of the interruptions and redirections will be undesirable. Tertiary emotions include such cases. Perhaps some addictions, obsessions, and some attentional disorders depend on transformations in the alarm mechanism.

Limitations of meta-management

Self-monitoring, self-evaluation and self-control are all fallible. No system can have full access to all its internal states and processes, on pain of infinite regress. Preferences used in selection and self-evaluation may be erroneous or ill-judged (e.g. self-evaluation based on religious indoctrination). Control over deliberative processes may be partial, e.g. because the global alarm mechanisms cannot safely be suppressed completely, or because of loud noises, addictions, etc.

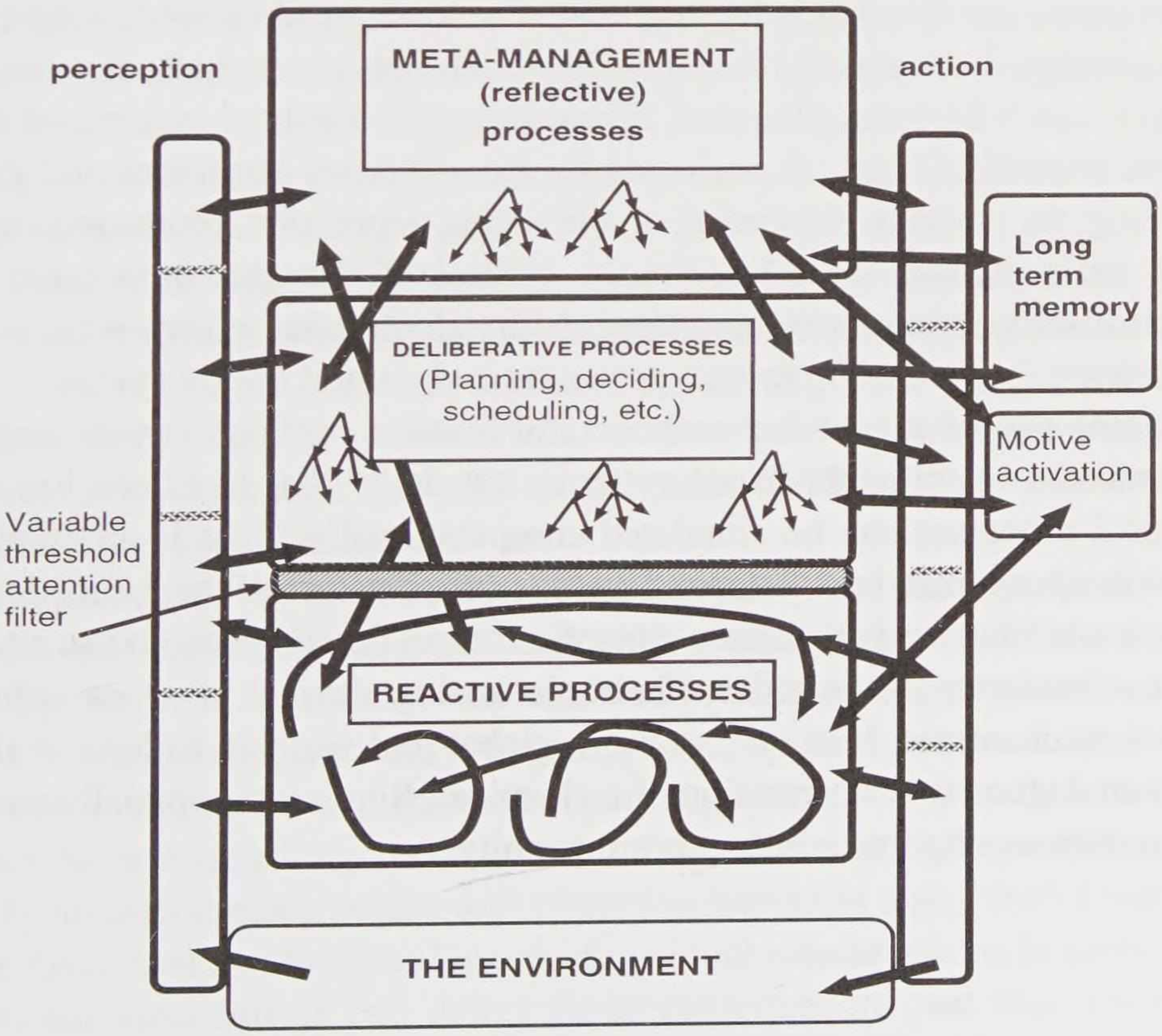


Figure 3. Towards a human-like architecture, with reactive, deliberative and meta-management mechanisms. (Alarm mechanisms not shown)

Figure 3 gives a crude indication of the sort of three layered architecture we are discussing, including showing (very inadequately) that perceptual and motor systems are also multi-layered. The alarm system is not shown because that would make the figure too cluttered (compare Figure 2). To envisage the addition of an alarm system in Figure 3 imagine an octopus on one side with tentacles extending into all the other sub-mechanisms, getting information and sending out global control signals.

6.4 *Non-semantic control*

The sorts of meta-management control sketched above involve precise direction of attention, or invocation of a strategy or evaluation of some state. These processes involve semantic content, e.g. reference to objects of attention or actions to perform. There is another type of control which produces global quantitative changes, for instance changing speed of operations, or degree of perseverance, or thresholds for attention diversion, or likelihood of adopting risky strategies. Some of these can be described as changes of *mood*. In animals a very complex chemical infrastructure is involved in some of these general control changes, as indicated by the effects of hormones and drugs, including, for instance, producing or alleviating depression, producing euphoria or hallucinations, reducing precise control of thoughts or actions. The chemical infrastructure can be influenced by alcohol and other drugs, smoking, disease, as well as by mental processes and natural bodily cycles.

Some people accept that architectural features outlined earlier could be implemented in computer-based systems, but doubt that functions based on chemical processes can be simulated computationally. That is an empirical question whose answer will depend on the precise nature of these functions. It may turn out that equivalent non-semantic control functions could use alternative mechanisms, for instance electronic analog devices or even software control mechanisms. It is easy to use a global real variable to have a global effect analogous to concentration of a chemical. Replicating spatially varying concentrations requires a little more ingenuity.

7. Some qualifications and implications

7.1 *Variability in meta-management*

Meta-management need not use a rigidly fixed collection of strategies. It should be modifiable by learning, so that one can detect new aspects of one's mental processing and evaluate them or control them in new ways. Examples would be learning to detect that one's grasp of a topic is confused, or that one is deliberating in a selfish way; and learning to disapprove of that sort of deliberation (which does not come naturally).

Instead of being rigid and monolithic, meta-management strategies may be different in different contexts. So the system may be thought of as 'occupied' by different 'control regimes' at different times, for instance: being a gentle parent at home, then driving a car aggressively, and becoming a cold and ruthless manager at the office.

Perhaps this is relevant to multiple personality disorders and other sorts of problems which lead people to seek therapy? This suggests many empirical questions. What are the 'role-switching' mechanisms? How can they go wrong? Can abuse in infancy produce long term damage in the architecture, and if so how?

Some of the states and processes described here, especially some of the high level emotional states in which there is a partial (and sometimes undesirable) loss of control, are not produced by mechanisms which evolved to produce them. They are side-effects, or emergent features of interactions between several mechanisms with other functions. Thus it is pointless asking what the functions of such states are.

In particular, the more sophisticated secondary and tertiary emotions are not directly implemented in an emotional mechanism, even if the simpler primary emotions are directly implemented in a global alarm system.

More generally, not everything supported by a mechanism is part of its function: multi-processing computer operating systems support thrashing, but do not have a thrashing mechanism! In some heavy load conditions they simply do far too much paging and swapping instead of doing useful work. So some functional mechanisms have dysfunctional consequences. In some cases additional mechanisms can detect those consequences and take corrective action, as in an operating system which detects that it is thrashing and prevents any new processes from starting up.

Sometimes a side-effect turns out to have beneficial consequences, which are then exploited for their effects. A person who finds that certain kinds of distress generate sympathy and support in others, may 'learn' to be distressed more often and in an exaggerated way. Likewise, a teacher may discover that real anger can be used to control a classroom, and learn to become angry (Second-order functionality).

7.2 *Forms of learning and development*

In such a complex architecture there are many different forms of development or learning that can occur, including: adding new capabilities to existing modules, creating new modules to extend the architecture, adding new links between modules, extending the formalisms used within a module (e.g. learning a new language, or a new notation for mathematical reasoning or music), storing new facts and associations in the long term factual memory, copying a new plan or strategy developed by the deliberative mechanism into the reactive mechanisms, thereby creating new reactive skills. In humans this kind of copying uses repetition of actions, with the deliberative system apparently supervising the training (or re-training) of the reactive system.

Different types of learning can be expected in different parts of the architecture. For example, perceptual mechanisms may learn to use new high level categories in classifying or interpreting perceived events. Examples are learning to read text or music fluently. In fluent reading the new percepts trigger *internal* actions. Action subsystems may learn to produce new complex orchestrated behaviours in response to more abstract "instructions" from the central mechanism.

Some new meta-management strategies, e.g. attention control strategies, seem to be produced by certain kinds of training, e.g. in meditation. The influence of a culture operating on the meta-management level can lead an individual to develop new ways of categorising and evaluating internal states, new forms of motivation, new motive generators, and new motive comparators, though much work needs to be done to explain how this works in detail. Cultural learning can vastly speed up learning by individuals. Forms of loving are also influenced by the culture, including disappointments caused by unrealistic culturally generated expectations.

Some subtle kinds of learning seem to involve the suppression of innate or previously learnt reactions. For instance emotional maturity includes learn-

ing to suppress or ignore some of the reactions of the global alarm system. This may include learning new strategies for adjusting the attention filter.

When a system can be changed in so many different ways, it is to be expected that in addition there are also many ways in which damage, disease, or genetic disability can change the system so as to interfere with its functioning. I suspect most professionals concerned with the identification and treatment of such problems, whether in education, counselling, psychiatry, etc. are aware of at most a tiny subset of the things that can happen. Perhaps these ideas will lead to helpful expansion of therapies.

7.3 *There is no unique architecture*

Many of the ideas sketched here are speculative. One problem is the difficulty of inferring architecture from known capabilities, since alternative architectures can in principle produce the same performance over a life time, as already remarked. Yet by analysing the trade-offs we may be able to rule out theoretically possible cases.

For instance, a purely reactive system could in principle do everything that can be done by a system with deliberative capabilities. However, the time required to evolve a collection of reactive behaviours large enough to cover the actions that a particular planning system could generate may be too long for the history of the universe. Moreover, storing them might require a brain too large to fit on the planet, and DNA molecules might be too small to encode them all. (Even the game tree for chess could not be fully encoded in any physical system.)

It may be that evolution 'discovered,' as AI designers have, that a good way to overcome these obstacles is to produce systems which are modular in the manner sketched above, and capable of explicit deliberation and planning. Or it may have found some alternative method which we have not yet thought of.

Unlike behaviourist psychologists and some AI researchers who reject explicit deliberation, I am inclined to regard human introspection, and everyday observation as providing at least *prima-facie* evidence for some of our capabilities. This gives me good reason to believe that people I know can plan many facets of a trip to an AI conference well in advance of taking a taxi to the airport. Likewise I know that people can memorise and (sometimes) reliably reproduce or use poems, stories, jokes, algebraic formulae, rules of many games, piano sonatas and moves in a dance. They can also do calculations and

problem-solving in their heads and report many of the steps. All this is evidence for the existence of some sort of symbol manipulating virtual machine, no matter how it may be implemented in brain mechanisms, and no matter what other mechanisms interact with it.

Even a besotted lover can dream about what he might have said during the last encounter, plan what he should do and say at the next one, and speculate about the thoughts and feelings of the object of his attention. 'What if' deliberative capabilities enriched by human language seem to be central to all aspects of human life, even if few other animals share them.

8. Conclusion

I have presented a collection of ideas some of which are very speculative while others are largely based on evidence gleaned over many years, including observation of a wide variety of humans of all ages, and what I have learnt from interactions with researchers from a range of disciplines including philosophy, computer science, biology, psychology and brain science, and reading their work. I have not tried to present all the evidence that inspired this work, since that would make the paper far too long, and in many cases I have not kept records. Many of the ideas are not original: much current research in AI involves investigating mechanisms of the kinds proposed here.

With colleagues and research students I am exploring some of the ideas (still in a very simplified form) in computational experiments using the Sim_agent toolkit, which runs under Poplog, and was specifically designed for such explorations. Code and documentation can be found at: ftp://ftp.cs.bham.ac.uk/pub/dist/poplog/

More detailed conceptual analysis, for which there is insufficient space here, would show that familiar mental states and processes such as seeing, deciding, wondering whether, hoping, regretting, enjoying, disliking, learning, planning and acting all involve various subtle and implicit sorts of information processing. (Many relevant ideas are in (Ortony, Clore & Collins 1988)).

Robots with meta-management capabilities allowing them to attend to internal virtual machine states, including intermediate sensory databases, might discover that they have qualia, and might wonder whether humans are zombies, since they are built quite differently. Work in progress explains this

in more detail and argues that when we understand the full nature of that information processing we shall see that it *suffices* to produce what we ordinarily understand by experience, consciousness, etc. (A draft is accessible via my web site, along with other papers elaborating on these ideas.)

Ultimately the ideas will need to be tested not on the basis of the evidence that suggested them, but on the basis of their explanatory power and ability to generate productive research. It takes time to distinguish what Lakatos referred to as 'progressive' and 'degenerative' research programmes, and there are no simple criteria of adequacy, for reasons I explained in chapter 2 of (Sloman 1978). There are still many unanswered questions, especially questions about the variety of information processing architectures, what their properties are, which ones could evolve naturally and which can only be produced by explicit engineering design. Investigating these questions requires collaboration between AI, Alife, Biology, Neuroscience, Psychology, Psychiatry, Anthropology, Linguistics, Philosophy, etc. Such work should not only be of scientific and philosophical interest, but may also lead to new developments in education, therapy and counselling. People often need professional help, but the professionals do not always understand normal functioning of the information processing architectures with which they are dealing, and therefore cannot account for failures and deviations from normality, nor provide help reliably except in a small subset of cases. A deeper understanding of information processing architectures and ways in which they can develop or go wrong could have profound practical significance.

Artificial agents may also need therapy and counselling, for the same reasons as humans. And existing human therapies may fail on them too!

Acknowledgements

This work has been supported by the UK Joint Council Initiative, The Renaissance Trust, and DERA Malvern. Many have helped with ideas, criticisms, software, and in some cases close collaboration, including: Steve Allen, Luc Beaudoin, Margaret Boden, Chris Complin, Monica Croucher, Kerstin Dautenhahn, Darryl Davis, Stan Franklin, Glyn Humphreys, Brian Logan, David Lodge, Roz Picard, Riccardo Poli, Louise Pryor, David Pym, Tim Read, Ed Shing, and Ian Wright. Anonymous referees were very helpful. Papers by various authors developing these ideas are in the Cognition and Affect Project ftp directory: ftp://ftp.cs.bham.ac.uk/pub/groups/cog affect/0-INDEX.html

References

Beaudoin, L. 1994. *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham.

Beaudoin, L. & A. Sloman. 1993. A study of motive processing and attention. In Sloman, A., Hogg, D., Humphreys, G., Partridge, D., & Ramsay, A. (Eds.), *Prospects for Artificial Intelligence*, pages 229–238. Amsterdam: IOS Press.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*.New York, Oxford: Oxford University Press.

Craik, K. 1943. *The Nature of Explanation*. London, New York: Cambridge University Press.

Damasio, A. R. 1994. *Descartes' Error, Emotion Reason and the Human Brain*.Grosset/Putnam Books.

Dennett, D. 1996. *Kinds of minds: towards an understanding of consciousness*. London: Weidenfeld and Nicholson.

Franklin, S. 1995. *Artificial Minds*. Cambridge, MA: Bradford Books, MIT Press.

Goleman, D. 1996. *Emotional Intelligence: Why It Can Matter More than IQ*. London: Bloomsbury Publishing.

LeDoux, J. E. 1996. *The Emotional Brain*. New York: Simon & Schuster.

Minsky, M. L. 1987. *The Society of Mind*. London: William Heinemann Ltd.

Mithen, S. 1996. *The Prehistory of the Mind*. London: Thames & Hudson.

Nilsson, N. J. 1994. Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research*, 1:139–158.

Oatley, K. & P. Johnson-Laird. 1987. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1:29–50.

Ortony, A., Clore, G., & A. Collins. 1988. *The Cognitive Structure of the Emotions*.New York: Cambridge University Press.

Peterson, D. (Ed.). 1996. *Forms of representation: an interdisciplinary theme for cognitive science*. Exeter, U.K.: Intellect Books.

Picard, R. 1997. *Affective Computing*. Cambridge, Mass, London, England: MIT Press.

Rose, S. 1993. *The Making of Memory*. Toronto, London, New York: Bantam Books.

Ryle, G. 1949. *The Concept of Mind*. Hutchinson.

Simon, H. A. 1967. Motivational and emotional controls of cognition. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

Sloman, A. 1978. *The Computer Revolution in Philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press).

Sloman, A. 1982. Towards a grammar of emotions. *New Universities Quarterly*, 36(3): 230–238.

Sloman, A. 1989. On designing a visual system (Towards a Gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337.

Sloman, A. 1992. Prolegomena to a theory of communication and affect. In Ortony, A., Slack, J., & Stock, O. (Eds.), *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, pages 229–260. Heidelberg, Germany: Springer.

Sloman, A. 1993. Prospects for AI as the general science of intelligence. In Sloman, A., Hogg, D., Humphreys, G., Partridge, D., & Ramsay, A. (Eds.), *Prospects for Artificial*

Intelligence, pages 1–10. Amsterdam: IOS Press.

Sloman, A. 1996a. Beyond turing equivalence. In Millican, P. & Clark, A. (Eds.), *Machines and Thought: The Legacy of Alan Turing (vol I)*, pages 179–219. Oxford: The Clarendon Press. (Presented at Turing90 Colloquium, Sussex University, April 1990. Also Cognitive Science technical report: CSRP-95-7).

Sloman, A. 1996b. *Towards a general theory of representations, in (Peterson, 1996)*, pages 118–140.

Sloman, A. 1997. What sort of control system is able to have a personality. In Trappl, R. & Petta, P. (Eds.), *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Berlin: Springer (Lecture Notes in AI).

Sloman, A. & M. Croucher. 1981. Why robots will have emotions. In *Proc 7th Int. Joint Conference on AI*, pages 197–202, Vancouver.

Weizenbaum, J. 1976. *Computer Power and Human Reason: From Judgement to Calculation*. W.H.Freeman.

Wiener, N. 1961. *Cybernetics: or Control and Communication in the Animal and the Machine*. Cambridge, Mass: The MIT Press. 2nd ed.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell. (2nd edition 1958).

Wright, I., A. Sloman & L. Beaudoin. 1996. Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126.

CHAPTER EIGHT

Connecting Reflection and Reaction

A Heterogeneous Multi-Agent Model

Ruth Aylett and David Barnes

1. Introduction

The problem that interested us in the work described in this chapter was how to combine two very different approaches to multiple agent activity. The first approach is the one taken by classical Artificial Intelligence and almost definitially seen there as 'intelligent' behaviour: explicit reasoning between alternatives using symbolic models of the world. The exemplar agent is human-like and is assumed to have an accurate and complete model of the environment. The environment in turn is assumed to be quite separate from the agent but passively available for the agent's activity, to which it responds in a completely predictable fashion.

Drawing out these assumptions about the nature of the agent and the environment in this way underlines the extent to which they may not be realistic, yet it should not be forgotten that this model of agent activity did not arise in a vacuum and does correspond to a particular type of human experience. Indeed, it has been argued that this is a characteristic human experience not shared by other animals — if only because humans have produced safe environments supporting prolonged reflection without immediate threats to survival. We call this the reflective approach.

The second approach, historically counterposed to the first from about the mid 1980s (for example Suchman 1987, Agre 1987, Brooks 1986), is charac-

terised by its view that agent and environment together form a composite and dynamic system (Agre 1987) and that agent activity is therefore always situated (Suchman 1987) and particular. If the first approach was centred on thinking, this is centred on interacting. The exemplar agent is often a social insect, such as an ant or a bee, and the environment is usually visualised as dynamic, containing as an important component other agents in mutual interaction with it.

In such a dynamic agent-environment system, modelling is seen as irrelevant if not impossible, with the coupling between the agent and the environment producing agent activity (Brooks 1991). In this approach, behavioural complexity — whether by an individual agent or a society of agents — is not produced by complex thought but emerges from the interaction of essentially simple components, the result, if anything, of a complex environment rather than a complex agent. We will call this the behavioural approach.

1.1 *Combining behavioural and reflective approaches*

Given the enormous difference between reflective and behavioural approaches, it may seem paradoxical or even senseless to wish to combine them. However there are in fact good reasons for wishing to do so. From a scientific perspective, the reflective approach to its human-like exemplar agent excluded perhaps the larger portion of human capability. Humans do not merely think — they also interact behaviourally and are coupled to their environment in the same type of way as any other animal. This would not matter if one saw reflection as a self-contained system independent of any others, but such a position is hard to maintain. From an evolutionary perspective, human behavioural capacities necessarily came first, and even at the level of individual development, it is possible to argue that cognitive processes arise through the internalisation of the patterns of social interaction (Vygotsky 1978), that is, through behavioural interaction with the environment.

A further scientific objection to treating reflective systems in isolation comes from the debate on symbol grounding (Harnad 1990). Searle, in his well-known 'Chinese Room' thought experiment, argued very strongly that the symbols in reflective systems had no inherent meaning but only acquired meaning in the heads of their designers. (Harnad 1990) also considered how the meanings of such meaningless symbols, manipulated solely on the basis of their (arbitrary) shapes, could be grounded in anything but other meaningless

symbols. His answer was that symbolic representations have to be grounded bottom-up in non symbolic representations. This suggests that a reflective system must have a behavioural system linked to it if its symbols are to be regarded as grounded in this sense.

For workers with real-world agents, that is, robots, there are also a number of very concrete reasons for exploring how reflective and behavioural systems can be combined. Experiments with robots soon demonstrate the limitations of agent architectures since the real world, even in a benign form such as the inside of a laboratory, is far more testing than a simulation. Architectures based on the reflective approach were found to be slow, fragile and extremely computationally demanding, to the point where a robot needed an umbilical to a bank of work stations to function, though not very quickly. The root of the problem (Brooks 1991) was the requirement to maintain a complete and accurate world model. Sensors do not deliver symbols, nor can they gather any data about parts of the environment out of their range. Thus world models are bound to be incomplete, are likely to be inaccurate and take a long time to build, during which time the real world may have moved into a different state. No worker in robotics wants their expensive robot to drive into a hole its world model has not recorded yet.

Behavioural architectures overcame these problems. Consisting of tight sensor-actuator couplings together with some conflict resolution mechanism, they are fast, responsive, make low computational demands, and will always try to do *something*. The assertion of the behavioural approach that complex behaviour can be produced by the interaction of simple components and with no modelling at all has been demonstrated by a number of groups (for example: Brooks 1986, Connell 1989, Genovese et al 1992, Kube & Zhang 1994) including ourselves (Barnes 1996).

An interesting point is that the emphasis on interaction means that the parallel organisation of multiple interacting behaviours inside one agent is not all that different in principle from the external interaction of multiple behavioural agents. It is very easy to think of a single behavioural agent as a society of mind (Minsky 1985), from which interesting agent behaviour emerges and in a similar way, to think of a society of behavioural agents as a superorganism from which socially interesting behaviour emerges without any explicit social knowledge in individual agents. As the discussion in the next section shows, we have demonstrated interesting and useful multi-robot cooperative behaviour using a purely behavioural model.

However the behavioural approach has itself a number of practical difficulties. A designed system that depends on interaction relies on the designer understanding what interactions will occur and what the effect will be. Once more than, say, a dozen behaviours are involved, this becomes almost impossible. Even where the number of interactions is tractable, not all of them may have the designer's intended effect. Thus obstacle avoidance behaviours, as is well known, interact negatively with any behaviour involving contact with an object, typically producing a deadlock, and even with navigation behaviours in certain configurations of obstacle and target location.

In reality, these are examples of the known tendency for local decision-making processes to settle into local rather than global minima in a search space. The behavioural approach tends to break down when success in an environment depends on a particular sequencing of discrete actions, not all of which, considered in isolation, contribute to the final goal. One might ask why, in this case, social insects are so successful. One answer is that such societies embody a huge amount of redundancy — the society will survive even if some proportion of the individuals are not successful at any given time or fail to survive themselves. This redundancy in turn hinges on a large-scale rate of reproduction of individuals. The economics of robots are very different.

A further concern is the *niche* (Hayes-Roth 1996) for which robots are normally designed. This is often a semi-structured environment constructed by humans in which a variety of specific tasks are to be carried out with reasonable efficiency. As 'members' of such a micro-society, robot-human cooperation may be as important — or more important — than robot-robot cooperation. Such cooperation is normally mediated by an organisational hierarchy so that robots are often asked to perform the role of operational staff supervised by a human manager. Purely behavioural systems present an opaque interface in such a social context and their overall functional inflexibility (they do what they do given the sensor input at the time) makes them capable only of a background service role — for example continuous cleaning.

1.2 *The MACTA project*

It was clear to us that the problems just discussed occurred in areas in which the reflective approach was particularly strong. In the project Multiple Automata for Complex Task Achievement, MACTA, we examined how a hybrid system could be created in which reflective and behavioural approaches were combined.

We noted that in particular, a symbolic-level AI planning system offers a straightforward human interface, through which a goal can be specified and decomposed into a sequence of actions. This sequence is such that if the actions are carried out successfully, the goal state will be reached without the agent getting stuck in a local minimum. The actions within the sequence, represented in partial order form — that is as a task network — give what might be called the subtask structure, in which it is clear what intermediate goals need to be achieved, thus providing a mechanism for dividing behaviours into more tractable groups. This parallels the way humans themselves organise complex tasks with explicit planning of a whole task and allocation of parts of it to various individuals to execute.

In order to avoid the major problem of the reflective approach in robotics — the attempt to build world models from sensor data — it was decided that the use of such a planner would be confined to a level of abstraction in which little changed in the agents' domain of activity. In the type of semi-structured environments just referred to — such as for example a factory — many features of the environments are relatively static, such as the number and position of rooms and the overall location of equipment and materials. Such a model need not be constructed in real-time from sensor data since it is composed mainly of *a priori* knowledge with a long shelf-life (Gat 1993). Therefore symbolic planning need not be mounted on robots at all but could reside on a fixed computer system, so long as it could be shown that only small amounts of interaction between the two types of system were needed. Again, we draw a parallel with human organisation, in which a supervisor is typically not directly involved in the execution of a task.

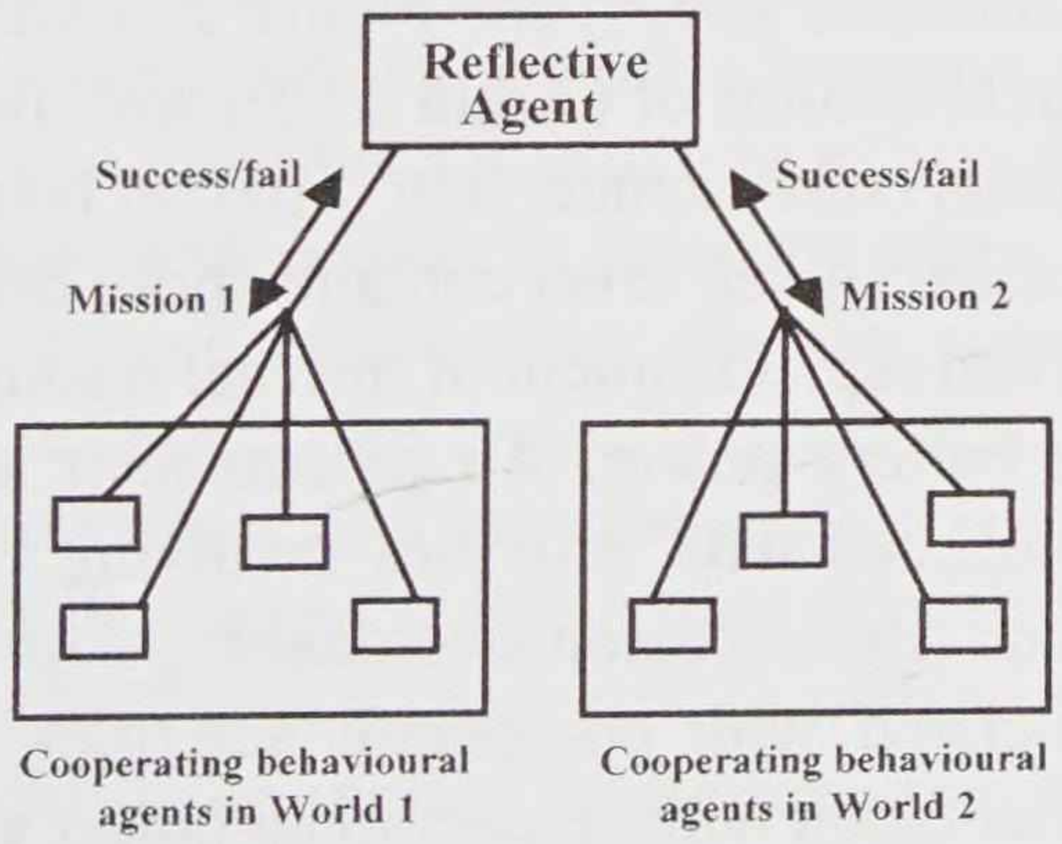


Figure 1. MACTA high-level architecture

In summary, a reflective agent could act at a supervisory level (Ephrati & Rosenschein 1992) to a number of behavioural agents with these agents using their behavioural architecture to cope with the variation and irregularity in the world as they encountered it. This arrangement maps very closely onto the organisational hierarchy already referred to. These considerations produced the overall architecture of Figure 1 in which the Reflective Agent was implemented on a fixed computer and the Behavioural Agents on robots both real and simulated.

In the rest of this chapter, we consider this architecture in more detail, starting with the Behavioural Agents, in the next section, and moving on to the Reflective Agent in the following section. We will follow this with an evaluation of the hybrid system, which forms a heterogeneous multi-agent architecture, and consider the issues and problems outstanding.

2. Cooperating Robots

The interaction of multiple behavioural robots can be regarded as a continuum between two diverse types of behaviour. At one extreme, the behaviour can be regarded as being egotistic, where a robot is concerned purely with self directed behaviour, e.g. energy conservation. At the other extreme their behaviour can be regarded as being altruistic, e.g. when a group of robots need to work together to perform some common task. Given the different behaviours that can be found in single robot and multi-robot scenarios, the research focused upon the design of a control architecture that could accommodate diverse and usually conflicting behaviour types. What emerged was the Behaviour Synthesis Architecture (BSA) see Figure 2, which complements the mobile robot control architectures of (Arkin 1989) and (Brooks 1986).

For purely conceptual convenience, four different behaviour levels in the architecture were identified: A *self* level contains those behaviours concerned with the maximisation and replenishment of internal resources, e.g. remaining stationary to conserve battery power. An *environment* level contains those motion behaviours associated with activities involving other objects within the robot's environment, e.g. collision avoidance. A *species* level contains those behaviours associated with co-operant activities e.g. maintaining a correct position and orientation with respect to an object while co-operatively relocating this object. A *task* level contains those behaviours specific to a

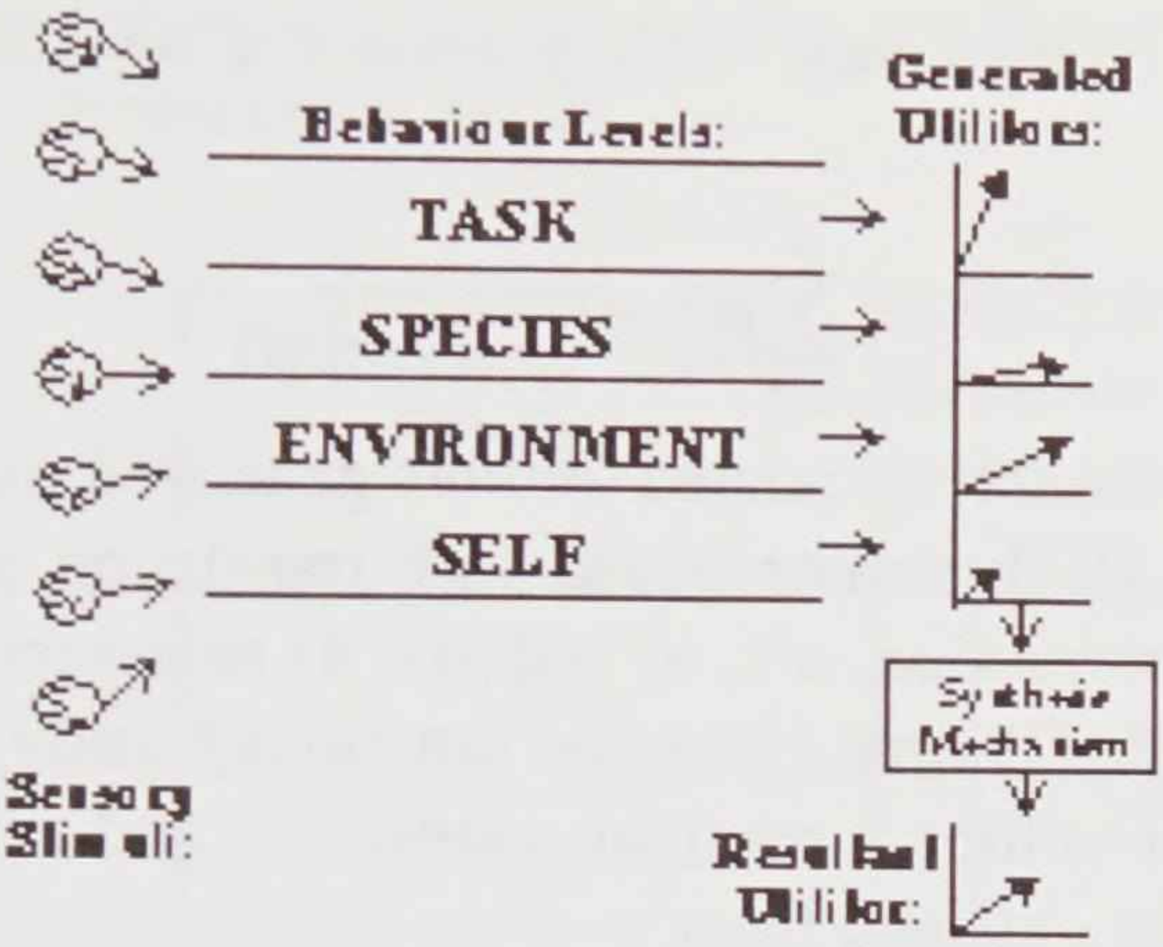


Figure 2. The Behaviour Synthesis Architecture

particular task, e.g. navigating to the initial location of an object to be relocated, then subsequent navigation to the desired goal location.

Sensory stimuli, from our developed robot sensor systems, provide the internal and external state information needed for the various levels and from each level, appropriate motion responses are generated that relate to the desired actuation. Any level can contain a number of *behaviour patterns*, **bp's**, where

$$bp = \begin{bmatrix} r \\ u \end{bmatrix} \text{and} \begin{matrix} r = f_r(s) \\ u = f_u(s) \end{matrix}$$

r is the desired motion response and is a function, f_r, of a given sensory stimulus, s. Associated to every response is a measure of its utility or importance, u. This quantity is a function, f_u, of the same sensory stimulus. Hence a **bp** defines not only what a robot's motion response should be for a given sensor input, but it also provides a measure as to how the relative importance of this response varies with respect to the same input. The values of r and u constitute a vector known as a *utilitor*. At any point in time, t, multiple conflicting motion responses are typically generated. For example, a robot may be navigating towards a goal location while co-operatively relocating an object when an obstacle unexpectedly appears in its path and at the same time it senses that it must re-charge its batteries. In such a situation, what should it do? In the BSA, conflicting motion responses are resolved by a behaviour synthesis mechanism to produce a resultant motion response. Competing utilitors are resolved by a process of linear superposition using the complex

exponential form for convenience. This generates a resultant utilitor, UX_t where:

$$UX_t = \sum_{n=1}^{m} u(t,n) \cdot e^{j \cdot r(t,n)}$$

and m equals the total number of related utilitors generated from the different behaviour levels, e.g. all those concerned with translation motion or those concerned with rotation motion. $u(t, n)$ and $r(t, n)$ represent the utility and response values generated at time t from the functions f_u and f_r which relate to **bp** n. Given a resultant utilitor, a resultant utility, uX_t, and a resultant motion response, rX_t are simply obtained from

$$uX_t = \frac{|UX_t|}{m} \text{ and } rX_t = \arg(UX_t)$$

X identifies the relevant degree of freedom, e.g. translate or rotate, and the resultant motion response, rX_t, is then executed by the robot.

Despite the success of the BSA, it did initially suffer from a problem common to all behavioural architectures. Effectively, **bp's** may interact in ways which are not useful to the robot. While utility functions are ideal in the context of generating a resultant robot motion, they are sensor dependent not sub-task dependent. Hence situations can arise when the associated utility for a particular **bp** needs to be forced to zero, irrespective of its input sensor value. This effectively produces a **bp** which does not contribute to the resultant motion response. We argue that the root of the problem is in allowing all **bp's** to be active at all times rather than restricting active behaviours to those most useful for the achievement of a particular sub-task. What was required was a means of allowing the task structure to create a context in which only appropriate **bp's** would be activated. In the BSA, a structure known as a *behaviour script* was designed for this purpose. A behaviour script consists of *behaviour packets*, each of which contain a triplet: (sensor pre-condition(s), **bp's**, sensor post-condition(s)). Sensor pre- and post-conditions are a combination of a particular sensor and either an initiating or terminating condition.

These are similar to the continuous action model implemented by (Gat 1992) in which activities are initiated and terminated by conditions, while Zelinsky's 'graphical sketches' (Zelinsky 1994) represent a more specialised form of this approach to navigation only. As each behaviour packet within the behaviour script is carried out, the pre-condition for the next is encountered so that finally, the whole script is executed. Hence this process constitutes an

ideal mechanism for sequencing behaviours, see Figure 3. Further BSA details can be found in (Barnes et al 1997).

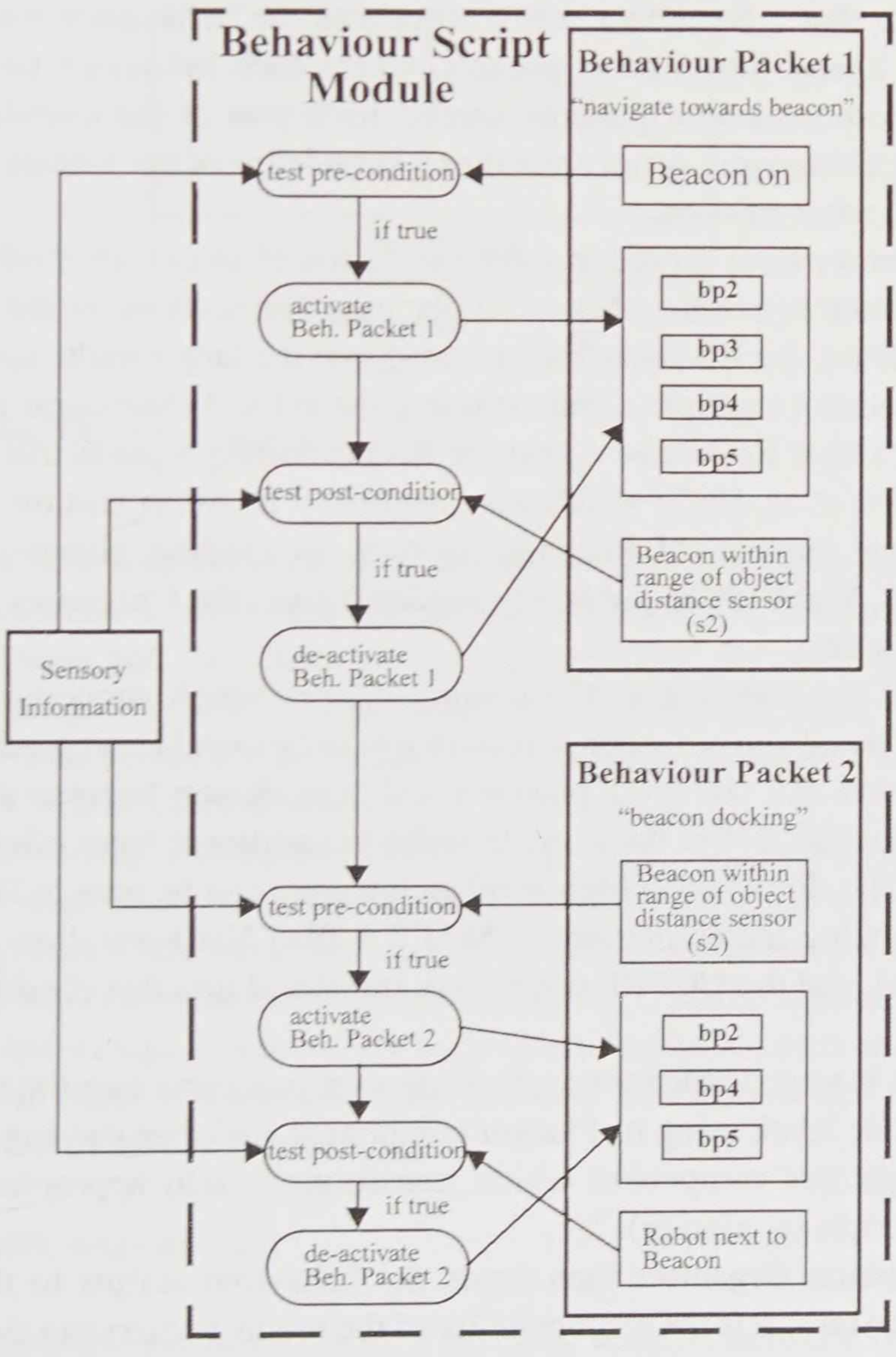


Figure 3. Behaviour script example. bp2–bp5 represent appropriate behaviour patterns while s2 is a robot to beacon distance measuring sensor

3. A Reflective Agent

Initially, the behaviour scripts discussed in the previous section were hand-crafted. Note that a behaviour script sequences the behaviours used by the behavioural agents into distinct packets, where each behaviour packet activates just those behaviour patterns needed for a part of the overall task. In other words, a behaviour script embodies a knowledge of the subtask structure of an overall robot mission.

However we have argued that the translation of an overall mission into a subtask structure is exactly what an AI planning system does. In the MACTA project therefore, the BSA was incorporated into the larger multi-agent architecture composing multiple agents running the BSA (behavioural agents, or BAs), and a single Reflective Agent (or RA) including a predictive planning system (Aylett et al. 1997) which was mentioned above in section 1. Rather than write yet another AI Planning System, an existing planner, UCPOP (Penberthy & Weld 1992), publically available from the University of Washington, was used.

UCPOP is a non-linear AI planning system which supports universal quantification and domain axioms, though it is in fact not hierarchical. It offers provable sound and complete planning and was chosen because it was the most suitable available at the time. In order to insulate it from other components in the RA, for which a high-level architecture can be seen in Figure 4, a standard planning representation — MACTA Plan Net Formalism (MPNF), was designed, and the UCPOP output was translated into this (Coddington & Aylett 1997).

The RA is responsible for transforming user goals into a partial order plan at the symbolic level, using its Planner component, and then passing this to its Mission Organiser component which transforms it into appropriate set of behaviour scripts (a *mission*).

The Mission Organiser then sends the behaviour scripts to the robots chosen for the task. It is the responsibility of the robots to carry out the mission autonomously, without further reference to the RA. Initially we have assumed successful missions, with each robot eventually returning a success message which is ultimately delivered back to the User Interface. However the problems of failing missions are under active investigation and mechanisms for coping with failure will need to be incorporated into reflective and behavioural agents in due course.



Figure 4. MACTA High-level Architecture

The relationship between reflective and behavioural components is intermediate between a traditional master-slave hierarchy in which one system decides actions and the others do exactly what they are told, and the more recent multi-agent systems in which each agent determines its own actions and negotiates with others (Davis & Smith 1983). The Agent Models component in Figure 4 will be discussed below.

In introducing a Reflective Agent in this way, a number of basic issues have to be confronted. One is the exact point at which activity ceases to be a planning problem and becomes entirely an issue for the behavioural agents. Another is how communication between the two types of agent is dealt with, and what knowledge is needed on each side to support communication. We will consider each of these issues in turn and then discuss the experiments carried out with the architecture.

3.1 *Dividing responsibility*

An AI Planning System produces a partially ordered network of actions in which actions that must follow each other are in sequence while those which could be executed in any order appear in parallel (see Figures 6 & 8 below as examples). Some planners — though not as it happens the one used in this implementation — are hierarchical, allowing more abstract actions to be expanded into more primitive ones. The most primitive actions used in a planner embody decisions about what is the responsibility of the planner, and

what that of the execution systems it is planning for, in this case behavioural robots.

For example, a planner primitive action *navigate-and-dock (agent, from, to)* makes the assumption that the BA can recognise both its start and end locations, that it can find the way between them and that it can dock at the end location. No planner help is supplied in these areas, and in particular, if a problem were to occur the execution agent would be expected to sort it out unaided. On the other hand, a planner could reformulate this action into a number of sub-actions: *navigate-segment (agent, intermediate1, intermediate2)* and *dock(agent, place)*. Here, the planner is contributing to navigation by planning a route between intermediate locations as well as specifically planning a single docking action.

Thus the practical determining factor in dividing responsibility is the behavioural capacity of the BAs (just as, in the human case, sensible supervisors base the detail of their instructions on the capabilities of the people who receive them). In our MACTA work, we chose primitive planner actions that mapped one-to-one onto behaviour packets in order to reflect this criterion. In this way, a *navigate-to-beacon* action in the planner maps onto a *navigate-to-beacon* behaviour packet within a behaviour script. The philosophy behind this approach was to put as much functionality as possible into the behavioural agents, and in consequence to raise planning to the most abstract level feasible.

A more interesting question, but one that is very difficult to answer, is what are the theoretical functional limits of a purely behavioural agent? Intuitively, the answer involves task complexity, however we choose to measure this. In particular, tasks with temporal complexity, in which a number of distinct intermediate states must be traversed correctly appear, as raised in section 1, to exceed the capabilities of the behavioural approach, unless we consider rigid sequences, as found in some insects. If one accepts that reflective abilities are not an all-or-nothing affair, and that a number of mammals employ reflection to one degree or another, then one way of approaching this question would be to consider the complexity of animal behaviour in relation to the reflective capacity of the animal.

3.2 *Knowledge requirements for communication*

The Behaviour Script represents the interface between the Reflective Agent and the Behavioural Agents in this system and defines the knowledge require-

ments of the BAs, which are minimal. The BAs contain no explicit models, in particular no environmental maps, cartesian or otherwise, and have no knowledge of the existence of other agents like themselves. Their knowledge is entirely procedural, embodied in the stimulus-response and stimulus-utility functions which make up the behaviour patterns in their repertoire. It is for this reason that negotiation between the BAs and the RA is not used: the BAs have no knowledge on the basis of which negotiation could occur. On the side of the Reflective Agent, the knowledge required for communication is much greater, and consists of whatever is needed to correctly translate plan networks into behaviour scripts.

As discussed in the previous section, the RA needs knowledge of the behavioural repertoires of the agents for which it is planning. Other non-behavioural attributes may also need to be known if agents are to be correctly selected for executing a particular plan. For example, if a robot is to be asked to carry an object, its weight capacity may be required. If a robot is to be asked to climb stairs, then its method of locomotion would be needed. This knowledge, together with 'contact' information — communication channel, protocol, and so on — is held in the Agent Models component of the RA architecture seen in Figure 4.

The Planner presents the Mission Organiser with a partially-ordered plan net containing primitive actions — that is actions which can be expanded no further by the planner — for conversion into behaviour scripts consisting of a series of behaviour packets. Each primitive action must have one or more agents allocated to it for which information is available in the Agent Models. The conversion process linearises the plan net into a totally ordered plan for each robot and then matches each primitive in the linearised plan for a robot to a behaviour packet in its behavioural repertoire, where each packet is indexed by symbolic versions of the pre- and post-conditions discussed in section 2 above.

This description of the action to packet conversion process as a matching of pre- and post-conditions must however be qualified since in fact not all planner goals are represented at the behaviour packet level — two very different roles are played by these constructs. Planner pre-conditions represent the *logically necessary* conditions for the correct behaviour to emerge when the corresponding behaviour packet is active; for example a multi-robot CARRY requires each robot to be holding the transported object. Planning is needed to make sure that these conditions will be met when the robots execute

the mission because the correct subtask structure has been communicated to them.

In contrast, at the execution level, *transition conditions* are important — the sensory inputs which deactivate one behaviour packet and activate the next and are derived from the context of the task. In one of the MACTA experiments, the *necessary* condition for a robot to release an object is that the robot is holding it, but the *transition* condition is that the appropriate beacon is sensed since the robots currently have no gripper sensor capable of telling them that they are holding an object. In the same way, the behaviour packet for a single-robot NAVIGATE is exactly the same as that for a single-robot CARRY since the robot has no sensor reporting the presence of the carried object unless it is collaborating with a second robot as in experiment 1 below.

3.3 *Planning for cooperative activity*

As can be seen from the experiments described below, a fundamental part of the MACTA project was the carrying out of actions involving more than one robot. We call these multi-robot actions. The first experiment described concerns the cooperative relocation of an object by two RWI B12 platforms, known in the project as Fred and Ginger, with one robot carrying each end. The second experiment includes the transfer of an object from one robot to the other and also the tracking of one robot by the other. We define cooperative multi-robot actions as one of two types. In *closely-coupled* cooperation, robots are physically connected. Thus when two robots jointly transport an object, the fact that each holds the object physically connects them and can be used as a form of communication. *Loosely-coupled* cooperation occurs when robots are not physically linked, for example as they are when tracking (and would be if moving in formation). More thought has to be put into the communicative requirements of loosely coupled cooperation.

Cooperative actions can also be categorised across an orthogonal axis depending on whether all robots involved in the action manifest the same behaviour (equal cooperation) or whether they play unequal parts in the action. While close- and loose-coupling affects the communicative requirements for the robots, *equal* and *unequal* cooperation has a basic effect on the behaviour scripts that must be produced, and is therefore far more of an issue at the planning level. For example, two robots can relocate an object with each trying independently to move it to the target beacon as in the first experiment

below — an example of equal cooperation. The planner primitive action designed for this is a multi-robot CARRY.

On the other hand, one robot could have the behaviour patterns necessary to get the object to the target beacon, and the other robot, only those behaviour patterns needed to go along with this. In this case the second robot could be seen as 'helping' the first one, rather than cooperating equally, and this distinction would have to be embodied in a separate planner primitive action HELP, for the correct behaviour script to be produced for each robot. The same type of distinction could be made between moving in formation, where each robot navigates independently, and herding, where one robot leads and all the others just follow. Our experiments to date all assume equal cooperation based on two identical real robots, but it is quite possible that robots might have different abilities and be unable to cooperate equally.

3.4 *Robot experiments*

The first phase of experiments concentrated on showing that the behavioural architecture of the BAs, the Behavioural Synthesis Architecture (BSA), provided enough competence for cooperative tasks to be executed by more than one robot. In the second phase, more complex tasks were tried in which the RA component of the overall system was more extensively tested. The robots used in these experiments were two RWI B12 platforms, known in the project as Fred and Ginger. In the second phase some experiments were also tried purely in simulation since this made it possible to examine tasks involving more than two robots.

3.4.1 *Cooperative relocation*

The purpose of this experiment was to demonstrate that robots could carry out close-coupled cooperation entirely by behavioural means using sensor feedback rather than models and explicit communication. Fred and Ginger were each equipped with a capture head — an instrumented spring mounted X-Y table installed on the top surface of each robot. Figure 5 shows the two robots carrying a pallet such that each end of the pallet rests on a robot capture head.

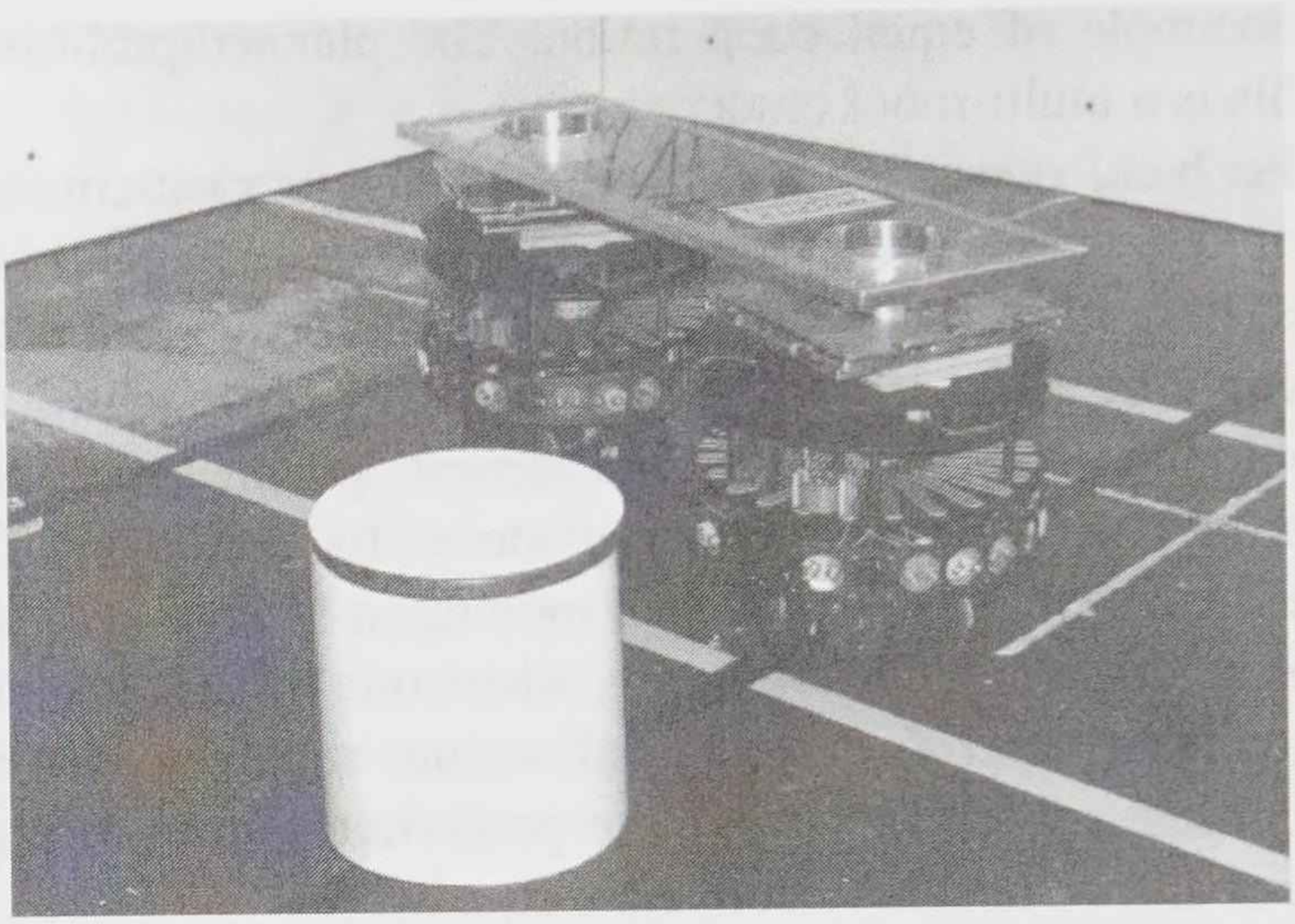


Figure 5. Experiment 1: Cooperative object relocation experiment

Each mobile robot contained the following behaviour pattern repertoire:

bp1 which ensures a steady translate (forwards) velocity.

bp2 which decelerates the robot when an object is detected to be within a given range.

bp3 which causes the robot to rotate away from an object when detected to be within a given range.

bp4 which ensures the robot translates to an external beacon destination in the shortest possible time, i.e. travels at a maximum given velocity.

bp5 which orientates the robot so that it is facing towards an external beacon destination point.

bp6 maintains a zero capture head velocity.

bp7 ensures that the position of the capture head remains central to its location on top of a mobile robot.

bp6 and **bp7** attempt to keep the capture heads centred on each robot, so that if one robot moves faster than the other, the capture head displacement has the effect of slowing the faster robot and speeding up the slower one. This is a bit like the sensory feedback two people carrying a table would receive from the pressure of the table and is of course a form of non-symbolic communication.

Figure 6 shows the simple $f_u(s)$ and $f_r(s)$ functions that were used for our experiments. The interaction between collision avoidance and capture-head

Description	Translate bp's
1. *translate always...* Generates a constant response with constant utility.	$r1 = 0.5$ $u1 = 0.5$
2. *obstacle avoid...* s2 - ultra sonic array. Decelerates robot when object detected. Negates any +ve bp's with large utility.	$r2 = \int_r(s2) = \sqrt{s2}$ $u2 = 0.95$
4. *translate to beacon...* s4 - beacon distance relative to robot. Beacon sensed at all times.	$r4 = 1.0$ $u4 = 0.05$
6. *centre capture head...* s5 - optical encoder, capture head velocity. Proportional response to sensor input with 'under-damped' utility.	$r6 = \int_r(s5) = s5$ $u6 = \int_u(s5) = \sqrt[4]{\quad (\mathrm{mod}\ s5)}$
7. *centre capture head...* s6 - optical encoder, capture head displacement. Proportional response to sensor input with 'over-damped' utility.	$r7 = \int_r(s6) = s6$ $u7 = \int_u(s6) = \quad (\mathrm{mod}\ s6)^2$
	Rotate bp's
3. *obstacle avoid...* s2 - ultra sonic array. Rotate away from object with large utility.	$r3 = \int_r(s2) = -s2 + 1$ $u3 = 0.95$
5. *rotate to beacon...* s3 - beacon bearing relative to robot. Beacon sensed at all times.	$r5 = \int_r(s3) = s3$ $u5 = \int_u(s3) = -s3 + 1$

Figure 6. The stimulus-response and stimulus-utility functions used

behaviours allows the robots to manoeuvre round the obstacles in the domain: as an object gets closer, collision avoidance gains an increased utility, and the nearer mobile robot will attempt to move away from it more strongly. This influences the motion of the other. Thus the system allows agents to communicate though they have no concept of communication and no exchange of symbolic information takes place. This is an example of the process sometimes known as stigmergy (Beckers et al. 1995) — or communication via the environment — which is often found in social insects, such as ants.

The complete behaviour script for each robot contains the following behaviour packets:

1. navigate to loading beacon
2. dock with loading station
3. cooperatively navigate to destination beacon
4. dock with unloading station.

Because packets activate behaviour patterns according to the sensor pre-and post-conditions within them, those behaviour patterns required for navigating to the docking site while avoiding obstacles will be active until the post-condition of the navigate to loading beacon pattern is met, (*distanceSensor: distance < threshold*). At this point the object avoidance behaviour **bp3** is deactivated, allowing docking to occur. In the same way, when the post-condition of the dock with loading station packet is met (*captureHeadSensor: object loaded*), the robots will move off, this time with **bp6** and **bp7** activated as well as the behaviour patterns for travelling to a beacon while avoiding obstacles. The condition (*captureHeadSensor: object loaded*) is in this case met by an active loading process, which itself waits until two robots have docked before loading the object onto their capture heads.

Figure 7 shows the robots' trajectories generated during the cooperative relocation experiments. The beacon is to the right of the figure and the robots were started one in front of the other, with both of them facing the beacon wall. Directly in front of the robots was an obstacle wall. The figure shows that the co-operating robots initially rotate so that each is directly facing the

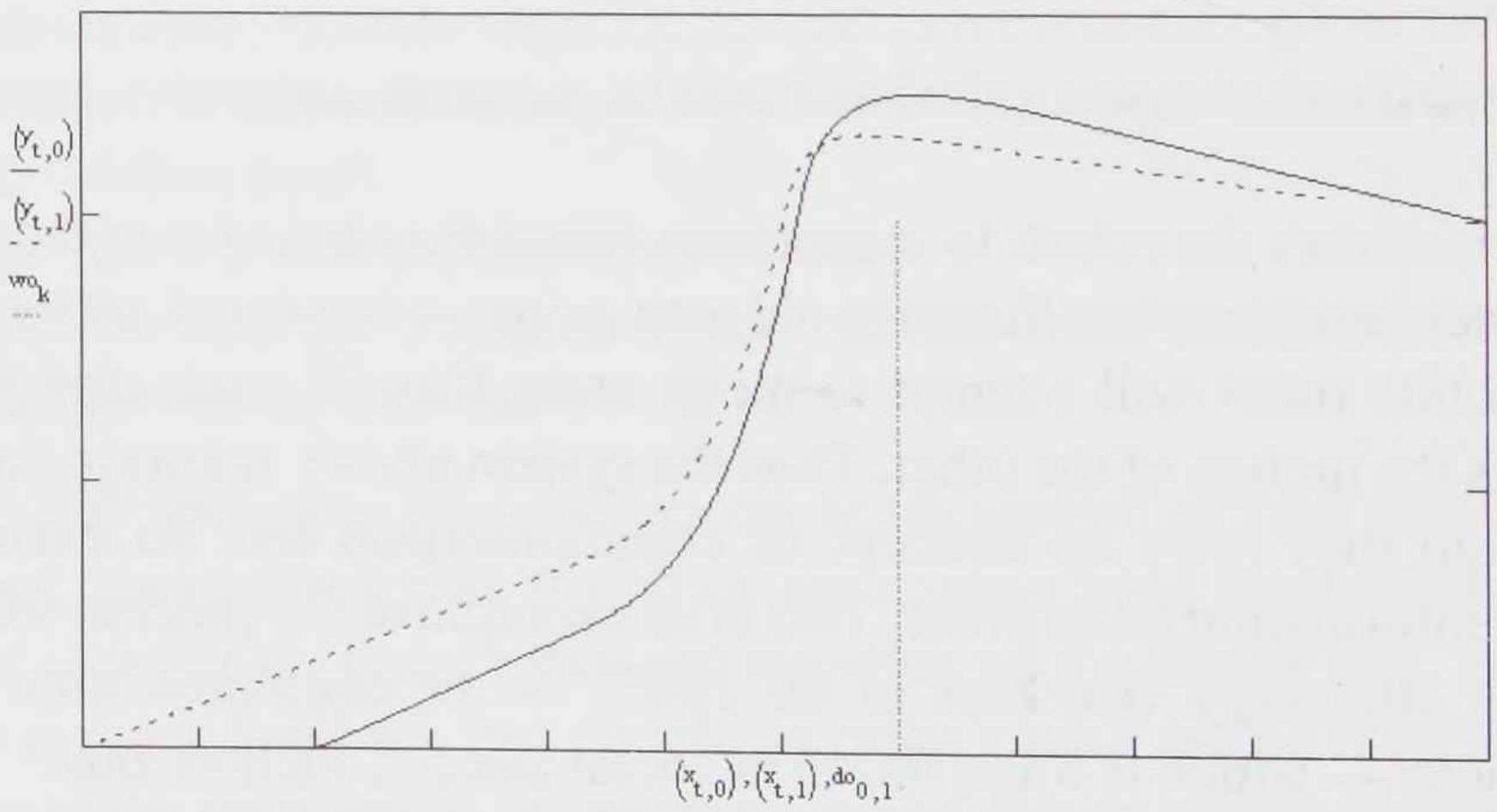


Figure 7. Robot trajectories during cooperative relocation

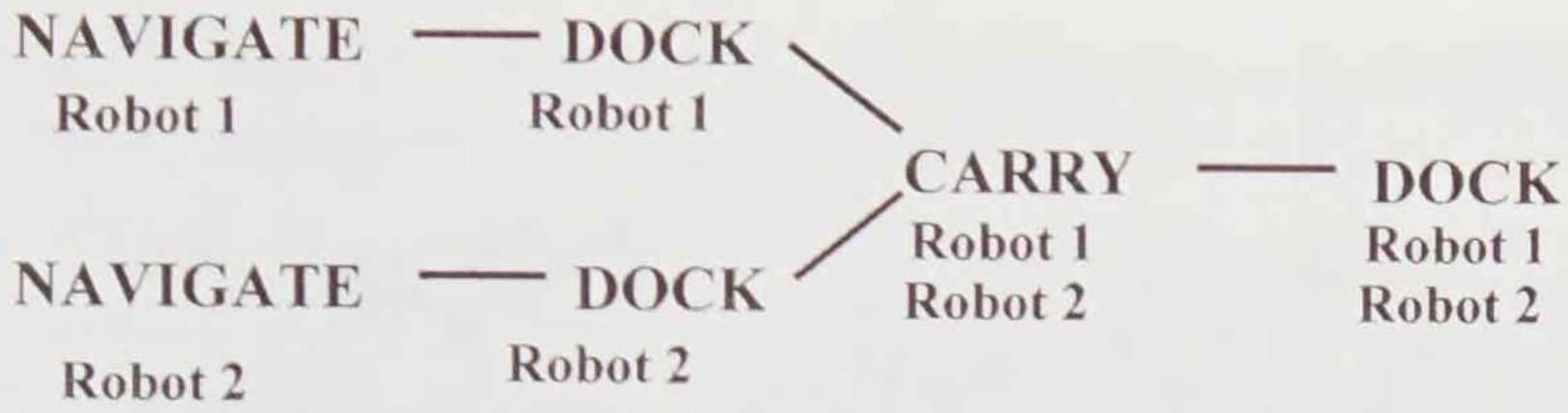


Figure 8. Plan net for object relocation

beacon. They then begin to translate towards the beacon but are forced to make a detour round the obstacle. Once past the wall they both reorientate themselves towards the beacon and finally translate up to the beacon prior to docking. The plan net for this task is trivial (Figure 8) and produces an identical script for each robot.

Runs of this experiment showed that cooperative relocation worked well as long as suitable utility functions are chosen for the behaviour patterns involved. This was done by finding approximate values in simulation and refining these with runs of the real robots. This trial-and-error procedure — while common in behavioural work because of the current lack of a rigorous analytical frame-work — is nevertheless far from satisfactory. Learning the necessary values would be much more desirable, but is itself a difficult task.

3.4.2 *Cooperative object relocation and tracking*

The aim of this experiment was to explore a task in which the robots had *different* behaviour scripts to follow and in which loose-coupled cooperation with no direct physical contact took place.

The robots Fred and Ginger were equipped with an arm carrying a gripper attached to the capture head and extensible in the X-axis — see Figure 9. In addition, a beacon was mounted onto Ginger. The robots started by cooperatively transporting a section of pipework, each robot using its gripper, to a destination beacon. When the beacon was detected, Fred released the pipe, and tracked Ginger as she transported it to a third beacon and docked with it. In a simulation experiment in this domain, rather than starting jointly holding the pipe, Fred handed it to Ginger — cooperative object exchange — and then tracked her as before.

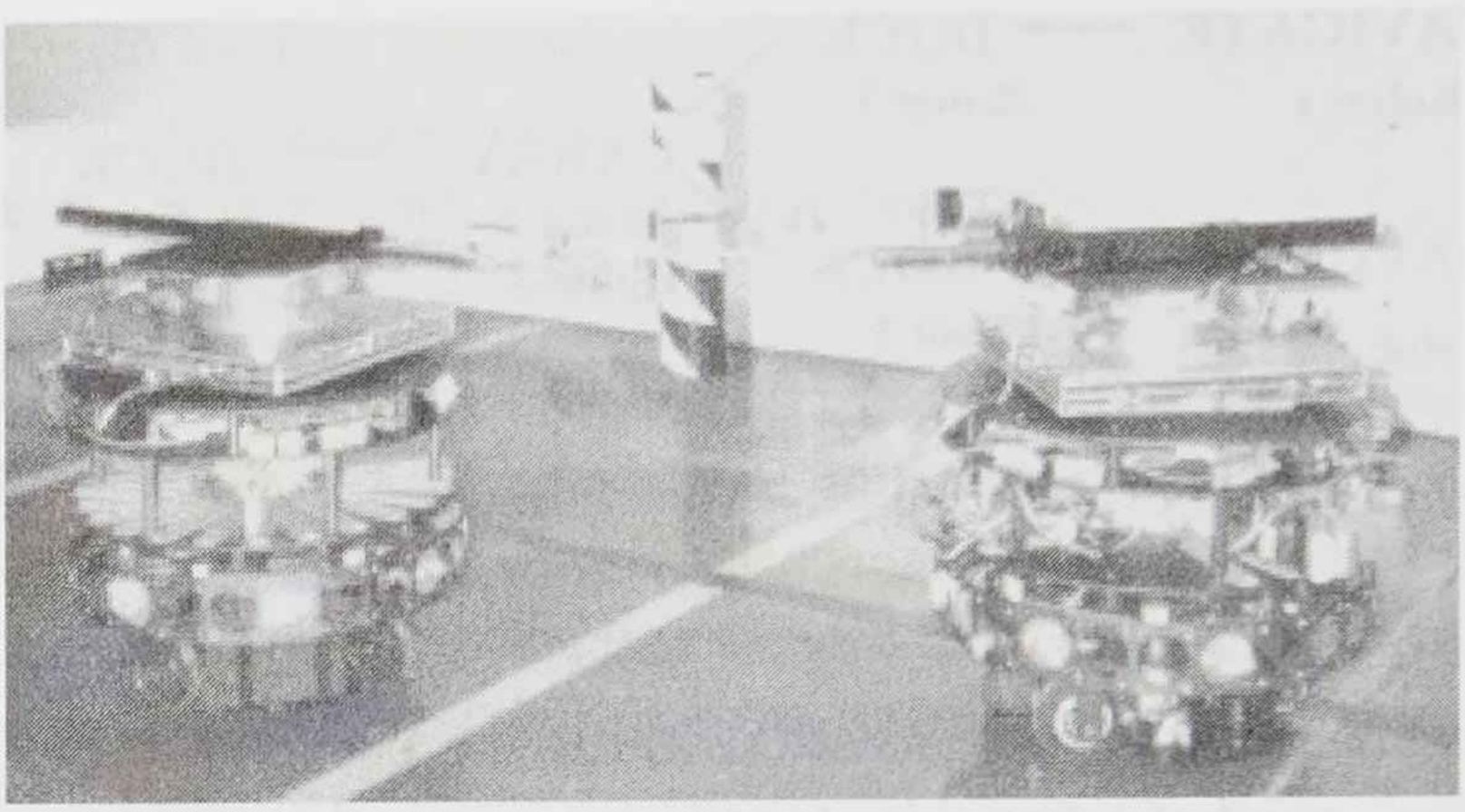


Figure 9. Experiment 2: cooperative object relocation and tracking

New behaviour patterns were developed to deal with the extension and retraction of the arm:

bp8 which extends the arm on detection of a beacon

bp9 which retracts the arm on detection of a beacon

Significantly, **bp1–bp7**, described above, turned out to be reusable. The two capture head behaviours allowed cooperative transportation even though arms were now in use and a totally different object was being transported. Tracking was produced by the existing navigate-to-beacon and object avoidance behaviour patterns since Fred was attracted by the beacon on Ginger but kept at a distance by the countervailing obstacle avoidance patterns.

An issue that arose in the conversion of the plan net into the necessary behaviour scripts was the instantiation of variables into the sensory pre- and post-conditions. The experiment involved three separate beacons, so that a navigate-to-beacon behaviour packet should have the beacon in question instantiated into it, identified, say, by frequency. Unfortunately, the infrared sensors on Fred and Ginger were not capable of making frequency distinctions between beacons so that this issue was explored in simulation only. The plan net for this experiment can be seen in Figure 10.

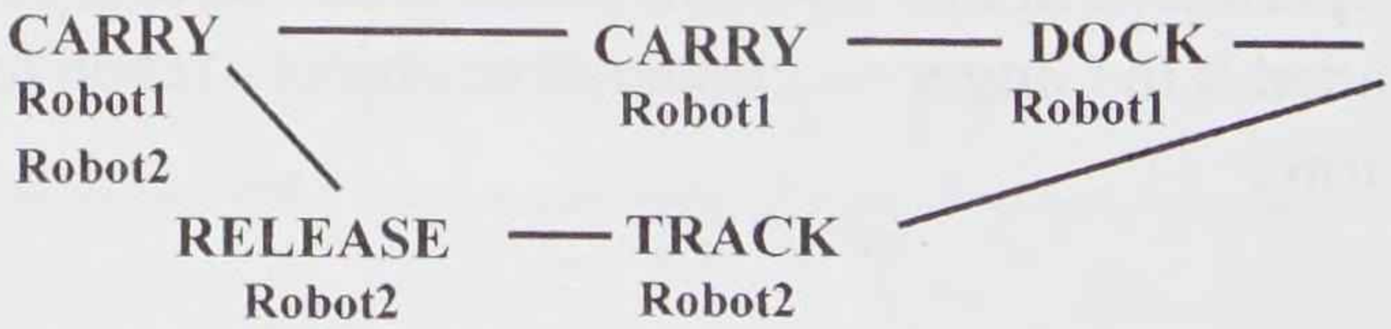


Figure 10. Plan net for second experiment

This experiment showed that it is possible to reuse behaviour *packets* at the planner level, especially if these packets are parametrised so that pre- and post-conditions can be extended to cover a range of situations. It also showed that the behaviour *patterns* within the packets could be reused in creating packets with a new functionality at the planning level. Finally, it demonstrates that loose-coupled cooperation can also be handled at the behavioural level as long as planning is used to produce a coherent sequence of packets and that the generation of different behaviour scripts for each robot requires no extension of the architecture.

3.4.3 *Simulator experiments*

In addition to the co-operant object relocation task, as performed by our two real robots, a sample store scenario was also chosen both for its relevance to real-world industrial problems of interest to the team, and as a way of investigating the effect of scaling-up a planning problem. To study the application of our hybrid reflective/reactive agent approach to this scenario, a simulator was implemented in C using PHIGS (Programmers Hierarchical Interactive Graphics System), a graphics library publicly available from the MIT X Consortium.

The sample store scenario was implemented in our MACTA simulator and is shown in Figure 11. This consists of two connected compartments. The upper area in the diagram is a sample store consisting of 12 storage locations divided by two aisles, while the lower area is entered by two doors.There are two interchange points between the two compartments (marked by two crosses). There are four robots (shown as unfilled circles with attached 'whiskers'), two in each compartment: each robot must remain within its allocated area. The two robots within the sample store are responsible for storing and retrieving samples (shown as filled circles) from their storage locations. Each of these robots is allocated an aisle.

An example task involving the storage of samples might consist of the following manoeuvres: a sample deposited at one of the doorways in the lower compartment is acquired by one of the robots within that area and taken to one of the interchange beacons situated between the two compartments. One of the robots in the sample store goes to the appropriate interchange beacon and acquires the sample prior to its storage at the desired location.

The plan-net actions and behaviour packets for this scenario were the same as those for the object relocation task described previously. However,

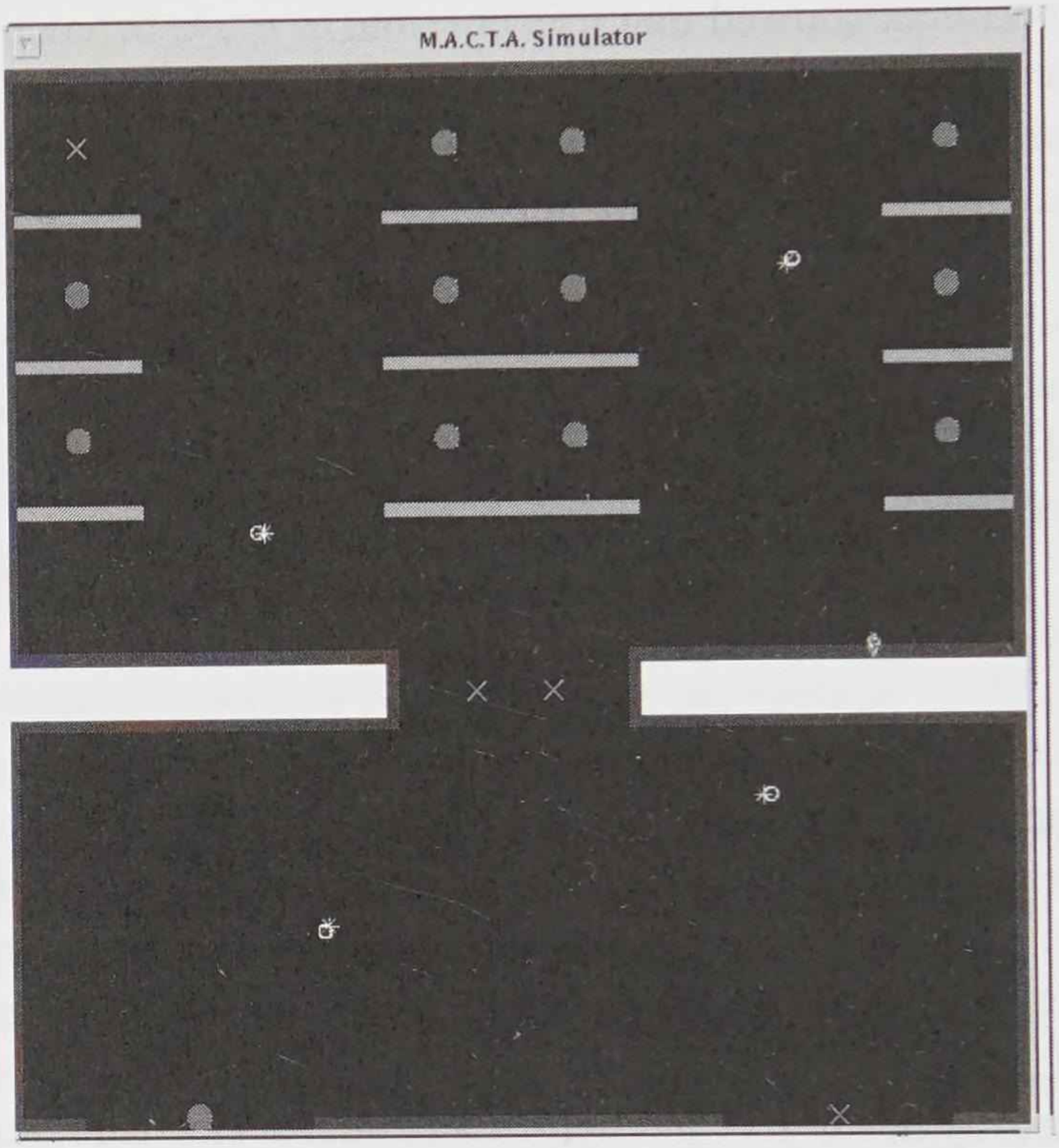


Figure 11. The MACTA simulator scenario

this scenario was considerably more complex from the point of view of the planner — there were four robots, each capable of sensing five beacon locations. Behavioural control of this scenario was a success, but it did expose the deficiencies of the UCPOP (University of Washington) planning system. UCPOP, as mentioned above, is a non-hierarchical planner and required excessive computation resources for this scenario. Nevertheless, the experiments undertaken using the simulator did demonstrate the ability of our MACTA architecture to deal with more complex multi-robot tasks than was previously achieved with just two real mobile robots.

4. Conclusions

The work in the MACTA project has been very successful in its own terms in that it has demonstrated that behavioural robots can carry out complex coop-

erative tasks if a reflective component is used to communicate the subtask structure to them. The RA acts as the 'consciousness' of the system in the sense that it has knowledge of the overall mission(s), and while not impacting upon the autonomy of the BAs, both defines the meaning of 'success' for the system and contributes towards the BAs arriving at it.

The robot architecture to which it is most similar is Connell's SSS (Servo, Subsumption, Symbolic), (Connell 1992). As in MACTA, a behavioural layer is supervised by a symbolic system, though in this case the subsumption architecture of Brooks is used and the symbolic level generates a *contingency table* which matches events with behavioural activations.This contingency table is effectively a reactive plan. However the architecture was applied to navigation rather than to a task with more complex sequencing, and was not extended to cooperating robots. It is not clear that the contingency table is as flexible as the MACTA equivalent — a behaviour script — or that it avoids the subsumption design problem of creating suitable networks of inhibition and exitation between behaviours.

A second comparison may be made with the 3T architecture (Bonasso 1995). This applies an AI planning system at its highest level, converting plans into Reactive Action Packets (RAPs) (Firby 1987) which manipulate a collection of soft, real-time routines called *skills*. However, while skills may be run in 'coherent sets of processes' (Firby 1996), they are not identical with behaviour patterns and emergent behaviour does not have the same dominant role as in a behavioural architecture. There is no general synthesis mechanism therefore in the way that there is in the BSA.

However it has been argued by a number of workers (for example Hayes-Roth 1996, Sloman 1999) that no single agent architecture is good for all purposes, just as no living creature architecture is good for all purposes. A good architecture is one that is adapted to its *niche*. Evaluating the MACTA architecture, therefore, means examining the niche for which it is intended as well as the architecture itself.

We have argued elsewhere (Aylett & Jones 1996) that planning domains can be characterised along the dimensions TASK, AGENT and ENVIRONMENT, together with the relationships between them. In this case, the MACTA architecture provides the content for AGENT, so that the niche can be thought of as a combination of the types of TASK for which the agent is thereby suited together with the ENVIRONMENT in which tasks are carried out.

A basic decision that must be made in a multi-agent system is whether

every AGENT will have the *same* predictive and reactive capacities or not, that is whether the system will be heterogeneous or homogenous. At one extreme, all agents could have the same abilities integrated in the same way, and there are a number of examples of this (Ferguson 1992, Muller & Pischel 1994). On the other hand, some multi-agent testbeds support the specialisation of agents, so that some agents are more reactive in their approach and others more deliberative (Cohen et al 1989, Grant & Lenting 1993, Farhoodi 1994).

The MACTA architecture has chosen to specialise agents and to concentrate planning in one place. The success of this choice rests on three things: the ability to separate the static from the dynamic aspects of ENVIRONMENT; to produce a TASK structure based only on the former; and to design behaviour patterns that can deal with the latter. This kind of separation may not be possible in all domains, but is likely to hold in the type of semi-structured industrial environments (factories, mines, docks, offices) targeted by the project.

A further constraint is embodied in the decision to sequentially chain behaviour packets into a behaviour script. This works well for domains in which the TASK has a definite structure, but would be inappropriate, for example, in robot football, where there is no pre-plannable sequence of subtasks. On the other hand, it would be straightforward to extend the architecture to allow behaviour packets to be defined in parallel at points in a behaviour script, so that a BA selected that packet from a group for which the sensory pre-condition had been met.

That the architecture is extensible in this type of direction has already been shown by its reimplementation in 'Virtual Teletubbies' (Aylett et al. 1998) — a virtual environment implementation of some BBC television children's characters. Unlike a television programme, a virtual environment need not have a sequential dramatic structure, so that the concept of a unitary behaviour script was replaced by that of 'sub-scripts' — smaller sequences of packets between which the virtual agent could switch according to its incoming sensory stimulus combined with internal state representing drives such as 'hunger' and 'boredom'.

A final constraint arises from the loose-coupling of the RA and BAs in this architecture. In order to give the BAs autonomy in the execution of tasks, very little communication takes place between RA and BAs. The behaviour script for each BA is communicated at the start of a task and a success or failure message is returned by a BA when the task is completed or when an

irretrievable error has occurred. This is appropriate for an ENVIRONMENT in which BAs may be geographically remote from the RA since it does not require high band-width communications nor does it require high grade reliability. However this makes the overall multi-agent system different from the usual conception of a 'society of mind' in which close-coupling and frequent interaction between agents is more usually the case.

While we referred above to the RA as the 'consciousness' of the overall system, in this respect it is very different from the human experience, in which abstract reasoning is integrated almost seamlessly with behaviours driven by sensory input. Consider, for example, driving a car, where abstract route planning and replanning is tied very closely to the semi-automatic behaviours required to operate the controls and perform short-range navigation. Arguably, human consciousness *inhabits* perception, as work investigating the dire effects of sensory deprivation would tend to suggest. We should note that in the driving example, human reflection does have access to an up-to-date real-time model of the state of the world derived from the sensors, and that it is this which makes the close-coupled relationship feasible and useful. It means that success or failure of goals at the reflective level can be detected very quickly so that replanning in real-time for particular problems can be applied. Without such accurate sensorderived models, which we have argued are currently not feasible for robots, close-coupling has little to offer.

References

Agre, P.E. & D. Chapman. 1987. Pengi: An implementation of a theory of activity. *Proceedings, AAAI-87*, Seattle, WA, 196–201.

Arkin, R.C. 1989. Motor schema-based mobile robot navigation. *Int. Journal of Robotics Research* 81(4), 92–112.

Arkin, R. 1998. *Behaviour-based Robotics*. The MIT Press.

Aylett, R.S. & S.D. Jones. 1996. Planner and Domain: Domain Configuration for a Task planner. *International Journal of Expert Systems*, v9 no 2, JAIR press, 279–318.

Aylett, R.S., A.M. Coddington, D.P. Barnes & R.A. Ghanea-Hercock. 1997. Supervising multiple cooperating mobile robots. *Proceedings, 1st International Conference on Autonomous Agents*, Marina Del Ray, 514515.

Aylett, R.S., A.J. Horrobin, J.J. O'Hare, A.A. Osman & M.M.T. Polshaw. 1998. Virtual Teletubbies: reapplying a robot architecture to virtual agents, *Proceedings, UKVRSIG-98*, in print. Available at http:// www.salford.ac.uk/iti/rsa/teletub.html.

Barnes, D.P. 1996. A Behavioural Synthesis Architecture for the Control of Mobile Robots. In: J.O. Gray & D.G. Caldwell (eds), *Advanced Robotics and Intelligent Machines*, IEE

Control Eng. Series 51, IEE London, 295–314.
Barnes, D.P., R.A. Ghanea-Hercock, R.S. Aylett & A.M. Coddington. 1997. Many hands make light work? An investigation into behaviourally controlled cooperant autonomous mobile robots. *Proceedings, 1st International Conference on Autonomous Agents*, Marina Del Ray, 413–420.
Beckers, R., O.E. Holland & J.L. Deneubourg. 1995. From Local Actions to Global Tasks: Stigmergy and Collective Robotics. In Brooks, R.A. & Maes, P. (eds), *Artificial life IV*, MIT press, 181–189.
Bonasso, R.D. 1993. Integrating Reaction Plans and Layered Competences Through Synchronous Control. *Robotica*, 1225–1231.
Brooks, R.A. 1986. A robust layered control system for a mobile robot. *IEEE Trans. Roobotics & Automation*, 2 (1986), 14–23.
Brooks, R.A. 1991. Intelligence without Representation. *Artificial Intelligence*, 47 (1–3), 139–160.
Coddington, A.M. & R.S. Aylett. 1997. Interfacing UCPOP to Multiple Execution Agents. *Proceedings, 16th Workshop of the UK Planning and Scheduling SIG*, University of Durham, 19–30.
Cohen, P.R., M.L. Greenberg, D.M. Hart & A.E. Howe. 1989. Trial by Fire: Understanding the Design Requirements for Agents in Complex Environments. *AI Magazine*, 10(3), 32–48.
Connell, J. 1989. A Colony Architecture for an Artificial Creature. *MIT AI Technical report* 1151.
Connell, J.H. 1992. SSS: A hybrid architecture applied to robot navigation. *Proceedings, IEEE International Conference on Robotics and Automation*, Nice, 2719–2724.
Davis, R. & R.G. Smith. 1983. Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence*, 20, 63–109.
Ephrati, E. & J.S. Rosenschein. 1992. Constrained Intelligent Action: Planning Under the Influence of a Master Agent. *Proceedings, 10th National Conference on Artificial Intelligence, AAAI 92*, 263–268.
Farhoodi, F. 1994. CADDIE: an Advanced Tool for Organisational Design and Process Modelling. In: K. Spurr, P. Layzell, L. Jennison & N. Richards (eds), *Software Assistance for Business Re-engineering*.
Firby, R.J. 1987. An investigation into reactive planning in complex domains. *Proceedings, International Joint Conference on AI, IJCAI '87*, 202–206.
Firby, R.J. 1996. Modularity issues in reactive planning. *Proceedings, 3rd International Conference on AI Planning*, AAAI Press, 78–85.
Gat, E. 1992. Integrating planning in a heyerogeneous asynchronous architecture for controlling real-world mobile robots. *Proceedings, 10th National Conf. on AI, AAAI-92*, 802–15.
Gat, E. 1993. On the Role of Stored Internal State in the Control of Autonomous Mobile Robots, *AI Magazine*, AAAI, 64–73.
Genovese, V., P. Daio, R. Magni & L. Odetti. 1992. Self-organising behaviour and swarm intelligence in a pack of minature mobile robots in search of pollutants. *Proceedings, 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 1575–1582.

Grant, T.J. & J.H.J. Lenting. 1993. An Arbitration Protocol for Inter-Agent Learning, Proceedings, *Proceedings, CKBS-SIG Workshop*, DAKE Centre, University of Keele, 185–209.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42: 335–46.

Hayes-Roth, B. 1996. An architecture for adaptive intelligent systems. In: P.E.Agre & S.J.Rosenschein (eds), *Computational Theories of Interaction and Agency*, MIT press, 329–66.

Kortenkamp, D., R.P. Bonasso & R. Murphy. 1998. *Artificial Intelligence and Mobile Robots*. AAAI Press/MIT Press.

Kube, C. & Hong Zhang. 1994. Collective Robots: From Social Insects to Robots. *Adaptive Behavior*, Vol 2, No 2 MIT, 189–218.

Minsky, M. 1985. *The Society of Mind*, Simon & Schuster 1985.

Penberthy, J.C. & D. Weld. 1992. UCPOP: A sound, complete, partial-order planner for ADL. *Proceedings, 3rd Int. Conf. on Priciples of Knowledge Representation and Reasoning*, 103–114.

Schank, R.C. & R.P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum.

Sloman, A. 1999. What sort of architecture is required for a human-like agent?, In: M.Wooldridge & A.Rao (eds), *Foundations of Rational Agency*, Kluwer Academic 1999, in print. Available at ftp:// ftp.cs.bham.ac.uk/pub/groups/cog_affect/Sloman. what.arch.ps.gz.

Steels, L. & P. Vogt. 1997. Grounding adaptive language games in robotic agents. *Proceedings of the 4th European Conference on Artificial Life*.

Suchman, L. 1987. *Plans and Situated Actions: The problem of Human-Machine Communication*. Cambridge University Press, UK.

Vygotsky, L.S. 1978. *Mind in Society: The Development of Higher Psychological Processes*, (eds) M.Cole, V.John-Steiner, S.Scribner, E.Souberman. Harvard University Press, (originally published in Russian in 1934).

Zelinsky, A., Y. Kuniyoshi & H. Tsukune. 1994. Monitoring and coordinating behaviours for purposive robot navigation. *Proceedings, Int. Conf. on Intelligent Robots and Systems*, Munich, vol. 2, 894–901.

Chapter Nine

The Role of Evaluation in Cognition and Social Interaction

Maria Miceli and Cristiano Castelfranchi

1. Introduction

This work attempts to provide an analysis of evaluation as a specific mental attitude, with a certain status and role within a knowledge representation system. We in fact believe evaluation is a central problem of knowledge representation, especially in goal-pursuing systems aiming at simulating intelligent autonomous agents, possibly in interaction with other agents.

Both artificial intelligence and cognitive and social psychology are recently paying special attention to naive and implicit theories of human behavior (Agre and Chapman 1990; Dweck 1991; Dweck 1996; Dweck and Leggett 1988; Epstein 1989; Hobbs and Moore 1985; Medin 1989; Murphy and Medin 1985; Pylyshyn 1984; Ross 1989; Wittenbrink, Gist, and Hilton 1993), commonsense reasoning (Anderson 1990; Doyle 1992; Evans, Newstead, and Byrne 1993; Galotti 1989; Kunda 1990; McDermott 1987; Oaksford and Chater 1991; Tverski and Kahneman 1974; Wason and Johnson-Laird 1972), and the relationship between cognition and motivation (e.g. Bratman 1987; Castelfranchi 1995; Higgins and Sorrentino 1990; Gollwitzer and Bargh 1996; Luck and D'Inverno 1995; Rao and Georgeff 1991; Simon 1979; Sloman 1987; Sorrentino and Higgins 1986). Those are all domains where evaluative processes play an important role. The explicit representation of evaluations should in fact be a crucial aspect in the study of planning and decision making, and any form of social interaction, both cooperative and competitive. However, though

evaluative processes are often implicit in the issues under study, evaluation seldom acquires the status of a topic. Abelson (1979) is one of the few who has explicitly addressed evaluation as a particular mental object; he himself says that his interest in evaluation remains "quite idiosyncratic", as it still seems to be. However, within the renewed field of the psychology of action, evaluations begin to receive some attention, for instance as far as their role in the various phases of goal selection and achievement is considered (e. g. Gollwitzer 1996).

Turning to the social domains, evaluation, which is often a necessary condition for the individual's decision to interact or not to interact, to propose or not to propose goal adoption and exchange, seldom receives direct attention. Often, it is just presupposed. Suffice it to consider the implicit part it plays in such classic themes as attitudes, prejudices, roles, and social representations.

This lack of focus on evaluation can be traced back to two main reasons: (a) on the one hand, the prevailing of poorly explicit representations of one's objects of study; in artificial intelligence consider for example the limitations of the rule based approach in expert systems (Steels 1987), or the ad hoc nature of operators in problem solving (Grosz and Sidner 1990). Also in social psychology, scarce attention to preliminary examination of the nature of the object of research is unfortunately quite common: rather than ask what evaluations and values are, the tendency is to set about sketching operational and in any case hazy definitions, and to begin *in medias res* with studies on, say, the values of adolescents, in which the said "values" come to overlap variously with other concepts, such as goals, needs and norms; (b) on the other hand, there is a more or less explicit assignment of evaluation to the domain of "hot cognition", with undue stress on its affective and emotional components. Undoubtedly, "affect" is often present in evaluation, but, first of all, it is not a necessary ingredient (there indeed exist also "cold" evaluations); secondly, the presence of affective components should not preclude the attempt at an explicit and formal treatment of evaluations and values (Miceli and Castelfranchi 1992).

1.1 *Aims of this work*

Our general aim is to identify evaluations and values as particular mental attitudes, and to explore their roles within a knowledge representation and goal-pursuing system. Of course, there exists a variety of knowledge represen-

tation and goal-pursuing systems, and consequently a variety of possible notions and roles of "evaluation" within those systems. So, first of all, it is worth specifying the meanings we assign to the basic notions on which we are building, such as agent, belief, goal, and action (for a more detailed treatment, see Conte and Castelfranchi 1995). This will allow to place evaluations and values in a more definite context, while showing why they can extend or improve the basic notions we are starting from.

To our current purposes, we restrict the notion of agent to that of *cognitive agent*. While an agent can be defined as a system that is able to produce some effects in the world by means of some regulatory mechanism, a cognitive agent is endowed with cognitive regulatory mechanisms, that is, symbolic representations such as goals and beliefs. So, a cognitive agent is an agent whose behavior is internally regulated by goals, which are in turn based on beliefs.

Our reference model of purposive behavior is cybernetic (e.g. Miller, Galanter, and Pribram 1960; Rosenblueth, Wiener, and Bigelow 1968). So, a cognitive agent's *goal* is a symbolic representation of the world that regulates (selects and monitors) its behavior, so as to adapt the world to that representation. (Actually, the basic mental object known as goal or regulatory state is the prototype of a complex family, including wishes, desires, intentions.) A *belief* is a symbolic representation of the world "as it is", i.e. a perceived state that the agent tends to adapt to the world (the prototype of the other fundamental mental object, including perceptions, opinions, knowledge, forecasts). As a consequence, an *action* (performed by a cognitive agent) is a pro-active, goal-oriented state transition (see Davidson 1996) which tends toward adjusting the world to the regulatory state, once a comparison is made between the latter and the perceived state. In fact, cognitive agents act because they record a discrepancy between their goals and their beliefs about the world.

As far as a formal (or pre-formal) treatment of such notions is concerned, we basically use Cohen and Levesque's (1990a, b) language, or rather a simplified version of it (see Conte and Castelfranchi 1995).

With that specified, let us now try to point out some basic reasons why evaluations, values, and related notions would be useful for improving any belief-desire-intention (BDI) architecture (e.g. Bratman, Israel, and Pollack 1988; Rao and Georgeff 1991).

We assume that a cognitive agent has several mutually independent end-goals (feeding, mating, being respected, etc.), a repertoire of elementary actions and the capacity to generate complex actions by combining several

elementary ones together; and that, given a goal, the agent can find subgoals and the sequence of actions required to attain the goal in a given situation. In order to do all this, a cognitive agent is guided by its beliefs (about the goals' attainability, means-end relationships, eliciting or favoring conditions, possible obstacles, and so on). Some specific belief structure is necessary to maintain and justify a given goal in a given state (for instance, a mere desire, or an intention, that is a goal chosen and planned for, or even in the process of being carried out). Without these supporting beliefs, the goal would be "dropped", or better, though it would not necessarily be terminated, its status would be changed (Castelfranchi 1996).

We assume that most of these supporting beliefs are evaluations of some kind. Evaluations are in fact beliefs about "what is good for what", and "why". Unlike mere "neutral" beliefs, they do not simply describe objects, states of the world or events and causal relations among them, but already put them into some means-end relationship with the agent's goals. So, evaluations already imply the assignment of a "value" to such objects (or states or events) relative to the agent's interest: things and events are "good" (hence, to be looked for, attained, or maintained) or "bad" (to be discarded, avoided, rejected); they are "powerful", or "powerless", or "dangerous" with regard to the agent's goals.

We suggest that the advantage offered by evaluations and evaluation-related notions is to make explicit what in most BDI architectures is just an implicit relationship — the relationship between beliefs and goals. Evaluations, as we view them, are in fact "bridge-notions" between beliefs and goals (see section 6.3). While in traditional BDI models that relation is to be found and established every time a goal arises, evaluations represent, so to say, already found and explicitly represented means-end relations, and solutions to problems (or classes of problems). The existence of explicit evaluations concerning the goodness of something for something else is a basic condition for goal-directed reasoning. Reasoning in turn allows either to justify one's behavior or to modify it, by changing the "values" assigned to the supposed means for one's goals.

In addition, as the cognitive agent we are interested in is a social one, we suggest that a special advantage is offered by evaluation and evaluative notions when dealing with a social agent. More precisely, we claim that, without those notions, it is quite impossible to construct so-called "socially situated agents". First of all, in fact, such notions allow to account for the fact that agents have some "value" for each other, in that, like objects and events,

they are viewed as good or bad, useful, useless, or dangerous, with regard to each other's goals; and the relationships that are established among agents depend on the value they attribute to each other. Secondly, since these values are attached by means of explicit beliefs about the qualities, skills, resources those agents are endowed with, such beliefs represent the starting point for social reasoning and for the social transmission of evaluative (both social and non social) knowledge. The social transmission of evaluations is in turn a powerful means for expanding, as well as modifying, through argumentation, persuasion and influencing, each other's evaluative equipment.

This general view justifies the need for constructing a set of meanings and relations relevant to the domain of evaluation — such notions as (social or non social) "power", "good", "bad", "useless", "dangerous", "enabling quality", "standard", and "value", which will be addressed in the paper.

In the following we first provide a basic definition of evaluation, distinguish between positive and negative evaluations (either of insufficiency or of harmfulness), and sketch a few types of different evaluative processes (sections 2–4). Second, we address the complex relationship between evaluation and emotion, trying, on the one hand, to distinguish between the two, and on the other to see how they are related, through the notion of emotional appraisal (section 5). Third, we move to the relation between evaluation and action, showing how evaluations are closely linked to goals (in that they are beliefs arising from goals, and that give rise to goals), and discussing the "bridge" function of evaluations: filling the gap between beliefs and goals; in so doing we address the role played by evaluation in problem solving and decision making (section 6). Fourth, we consider the role played by evaluation in social behavior, and point to the evaluative abilities an agent should be endowed with to act in a social context (section 7). Fifth, we address the notion of value: we define it as a special kind of evaluation, in that, while the latter implies a means-end relationship (something is assumed to be good or bad for a goal *p*) a value is a "means" for an unspecified goal, or class of goals, and turns into something "good in itself"; we analyze the relationship between value and related concepts, in particular goals and norms; we try to account for some typical features of values — their being unfalsifiable, indefinite, normative, and terminal; and we address the functions, both cognitive and social, of values, showing how those functions result from the borderline nature of values, between evaluations and goals: in particular, the social function is traced back to values' kinship with absolute imperatives, while the cognitive

functions are put in relation with values' evaluative features (sections 8 and 9). Finally we provide some concluding remarks (section 10).

2. What are Evaluations

We define an evaluation of an entity *x* as *a belief of an evaluating agent e about x's usefulness with regard to a goal p*. If for instance *x* is a pair of scissors and *e* believes — from direct experience, inference, or someone else's communication — that it is good for cutting a piece of cloth, in so doing *e* is evaluating the scissors with regard to that goal. We might represent an evaluation as follows:

(1) (BEL e (GOOD-FOR x p))

where *x* denotes an entity variable (i.e., an object of any kind: physical object, organism, agent, etc.), *e* is an agent variable and *p* is a well-formed formula representing a state of the world; the predicate (GOOD-FOR x p) means that *x* is a means for *p*, which is what *e* BELIEVE.

To be more precise, since *e* actually evaluates a certain *use of x* as GOOD-FOR *p*, we should express an evaluation as

(2) (BEL e (GOOD-FOR q p))

where *q* stands for the particular use of *x* that is viewed as a means for *p*. In other words, *q* will be equivalent to (USE a x), where USE denotes some form of exploitation of the entity *x*, and *a* denotes an agent variable, i.e. the "user" (that may happen to coincide with the evaluator *e*, but can also be different from *e*). If compared with (1), (2) also offers the advantage of being more general, in that it easily covers all those cases where the object of the evaluation is a state or event rather than a certain entity. For instance, *e* might evaluate "being fair-haired" as good for a woman to attract a man; or "playing the drums" as good for having fun; and so on. However, since in most of this paper we will consider evaluated *entities*, just for the sake of simplicity here we will take (1) as our reference definition, unless otherwise specified.

GOOD-FOR has a very broad meaning: it merely expresses that *x* (or *q*) is useful for making *p* true; *x* may either directly realize *p*, or cause *p* to be realized, or favor that *p* be realized. GOOD-FOR implies any of the following descriptions of the relationship between *x* and *p*: *x* may cause, produce, enable,

favor *p* to be true. In addition, a special restriction holds on the second argument of GOOD-FOR: *p* is not only a possible state of the world, but a goal, i.e., a state or event that somebody wants realized. Therefore, a broad definition of GOOD-FOR, which takes the predicate FAVOR to express the loosest possible relation between *x* and *p*, might be:

(3) (GOOD-FOR x p) $=^{def}$ (FAVOR x p) $\wedge$ (GOAL a p)

where (FAVOR x p) means that *x* favors that *p* is realized, *a* denotes an agent variable, and (GOAL a p) means that *a* has the goal that *p* is realized.

A number of specifications are needed. First of all, consider that we are taking a notion of *positive* evaluation as our starting point; when dealing with negative evaluations (section 3), we will need to take other aspects into account.

Second, *a*, the holder of the goal, might coincide with *e* when an evaluation is produced. It often happens, in fact, that one evaluates something with regard to a goal of one's own. But this is not necessary: *e* may evaluate *x* with regard to someone else's goal (suppose *e* is trying to solve a problem for *a*, and is looking for an adequate means for *a*'s goal). Indeed, *e* may even produce evaluations of the hypothetical kind, about possible goals of hypothetical agents, and reason as follows: *if p* were a goal in somebody's (*a*'s) mind, *x* would be useful for *p*.

As already observed, GOOD-FOR expresses a very broad means-end relationship. We believe it necessary to represent such a broad meaning, since, according to natural languages and commonsense, a very general and abstract representation of the means-end relationship looks anything but unusual. Think for instance of the abstract meaning of the word *good*, which can be attributed either to a knife or to a tree, or to a dentist, or to a friend, etc. Anyway, an analysis of possible sub-types of GOOD-FOR relation is of course worthwhile. In other words, GOOD-FOR is necessary but not sufficient: it is the top level of an abstraction conceptual hierarchy that presents, at its lower levels, more concrete and specific kinds of "power", like skills, qualities, and external resources (Miceli and Castelfranchi 1992); but we do not dwell on this topic here.

3. Kinds of Evaluation: Positive Versus Negative

An evaluation is *positive* when *e* believes that *x* favors the achievement of *p*, i.e., it is good for that goal, as in (1).

At first sight, then, an evaluation would be *negative* when *e* believes *x* (for instance, our scissors) is insufficient or inadequate as a means for *p* (to cut a piece of cloth), that is it does not favor *p* to be realized:

(4) (BEL e ¬(GOOD-FOR x p))

However, this is just one kind of negative evaluation, that of *insufficiency* or *inadequacy*. In fact, another kind of negative evaluation exists, where *e* believes that *x* is endowed with negative power, i.e. power to *frustrate* a certain goal *p*. Suppose *p* is the goal of safety, and our scissors are evaluated as *too* sharp, that is, they can hurt someone; such a negative evaluation of *harmfulness* or *dangerousness* might be expressed as follows:

(5) (BEL e (BAD-FOR x p))

where the predicate BAD-FOR is defined as:

(6) (BAD-FOR x p) $=^{def}$ (FAVOR x ¬p) ∧ (GOAL a p)

that is, *x* favors the realization of a certain state of the world ¬*p*, which is the opposite of someone's goal *p*. This kind of representation allows to express an important feature of any negative evaluation of dangerousness: the fact that *x* is still viewed as a means for something, albeit of a negative sign. Unlike what happens in the negative evaluation of inadequacy, here *x* is considered as "powerful", i.e. able to produce or allow some result. In a sense, a negative evaluation of dangerousness potentially implies some positive evaluation: it is sufficient to conceive ¬*p* as a possible goal in someone's mind! In that case, *x* would be viewed at the same time as both good for ¬*p* and bad for *p*, which is anything but unusual: a dangerous pair of scissors is in fact a *good* means for hurting people.

As one can see, the two kinds of negative evaluation we have just defined are very different from each other, as powerlessness is different from any form of negative power. The attitude and behavior of the evaluator with regard to *x* will in fact depend on the kind of negative evaluation. No doubt, a general attitude of avoidance is at stake in either case; but, if *x* is viewed as inadequate, *e* is likely to discard it, while if it is dangerous or harmful, *e* is likely to defend against it. This is particularly apparent when the evaluation concerns goal-

pursuing agents, rather than simple objects: while an evaluation of inadequacy affects the domain of "competence" of *x*'s image, an evaluation of harmfulness affects the moral domain of responsibility (Miceli 1992). Not surprisingly, *e* will show quite different behaviors and emotional reactions (pity or contempt toward the inadequate *x*; anger or fear toward the harmful *x*). But we will consider such issues as the role of evaluation in social behavior in section 7.

4. Kinds of Evaluative Process

"Evaluation" is an ambiguous term, in that it refers both to the cognitive process of evaluating and to the result of such a process. So far, we have analyzed some features of evaluation as a mental object. Let us outline now the most common processes through which evaluations are produced. We propose at least three general kinds of evaluative processes, provided they are not necessarily mutually exclusive (that is, more than one process might occur in a given context).

4.1 *Evaluation by performance*

Here *e* observes the "performance" of *x* with regard to *p*; that is, *e* tests *x*'s usefulness with regard to a given goal. For instance, *e* uses these scissors, trying to cut a piece of cloth, and, on the grounds of the result obtained, comes to assume whether (and to what degree) they are good for that goal. In other words, the evaluator, to answer the question "is *x* good for *p*", tries to empirically verify whether, once given *p* as a goal, a FAVOR relationship holds between *x* and *p*. In a sense, *e* decomposes the predicate GOOD-FOR into its basic constituents (see (3), the definition of GOOD-FOR), and assesses if the latter are true, as regards the given *x* and *p*. This looks like the more basic kind of evaluative process, and also the more reliable: after all, the proof of the pudding is in the eating!

4.2 *Evaluation by classification*

Evaluations behave like any other item of knowledge. So a general cognitive property such as inheritance also applies to evaluations: a new entity can inherit an evaluation from the class it belongs to.

In evaluation by classification, the evaluator's starting point is quite different from that of evaluation by performance. Here *e* already possesses a general evaluation about a class of objects (or states or events); that is, *e* already believes that *X* (say, scissors in general) are GOOD-FOR *p*. So, when faced with a specific *x*, *e* tries to assess whether that *x* belongs to *X*, i.e. it is as an instantiation of the class of *X*, already known as GOOD-FOR *p*. In other words, the starting question, "is *x* good for *p*", is translated into "does *x* belong to *X* (that are good for *p*)". If it is so, *e* infers that *x* is GOOD-FOR *p*.[1]

Therefore an evaluation by classification of an individual entity *x* with regard to *p* will result from an inferential rule of the following kind:

(7) $(\text{BEL e } ((\text{GOOD-FOR X p}) \wedge (\text{ISA x X}))) \Rightarrow (\text{BEL e } (\text{GOOD-FOR x p}))$

4.3 *Evaluation by standard*

Also in this case *x* is seen as an instantiation of the class of *X*, which has been already evaluated as GOOD-FOR *p*. But, in addition to the mere evaluation by classification, here *e* knows *why X* are GOOD-FOR *p*, that is, *e is explicitly representing those qualities of X which enable this class of entities to be* GOOD-FOR *p*. We call *standard of X* relative to *p* the set of *qualities which X should necessarily possess in order to make p true*. In the case of scissors, for instance, a certain degree of sharpness, certain dimensions, the kind of material they are made of, a certain weight, and so on will constitute their standard with regard to the goal of cutting a piece of cloth. If scissors are endowed with those qualities, they are GOOD-FOR *p*.

A standard might be expressed as follows:

(8) $(\text{STAND Q X p}) =^{def} (\text{QX} \supset (\text{GOOD-FOR X p}))$

where *Q* denotes the set of enabling qualities, and *QX* means that *X* possesses *Q*. So, *Q* is the standard of *X* with regard to *p* if *X*'s possessing *Q* implies its being GOOD-FOR *p*.

An *evaluation by standard* of an individual entity *x* with regard to *p* will result from an inferential rule of the following kind:

(9) $(\text{BEL e } ((\text{STAND Q X p}) \wedge (\text{ISA x X}) \wedge \text{Qx})) \Rightarrow (\text{BEL e } (\text{GOOD-FOR x p}))$

That is: *e*'s beliefs that a certain standard *Q* applies to *X* for *p*, that *x* belongs to *X*, and that *x* conforms to *Q* make *e* infer that *x* is GOOD-FOR *p*.

So, when a standard is available, *e*'s evaluative work with regard to *x*

consists in comparing its qualities with those included in the standard for its class, to see whether *Qx* holds, i.e., *x* conforms to the standard. In other words, the starting question, "is *x* good for *p*", is translated into "does *x* possess *Q*".

The construction of a standard implies remarkable cognitive capacities of explanation and generalization. First of all, building a standard implies not only assessing that *x* is good (or bad) for *p* (as in evaluation by performance or by classification), but also establishing *why* it is so. In fact, often we are not satisfied with the simple belief that *x* is GOOD-FOR *p*: we want to know why this happens and so try to find out the specific qualities that make *x*'s performance possible (for instance, we realize that it is these scissors' sharpness, heaviness, etc. which enable them to cut the cloth). Second, once the enabling qualities have been identified, *e* must make some generalization from *x* and its qualities to the class of *X*. At the end of this process, *e* will be able to conclude that, to cut a piece of cloth, scissors in general should present a certain degree of sharpness, heaviness, etc.

As one can see, standards radically modify the evaluative activity. When standards are available, evaluation can do without either trial and error or analogical procedures: *e* will not have to empirically test a given *x* with respect to *p*; nor will *e* have to see whether, more or less vaguely, *x* resembles a certain *x'* or belongs to a certain class of *X* which used to be good for *p*. On the contrary, *e* will analytically compare the specific *x* with its class, to see whether *x* possesses the required qualities described in the standard.

5. Evaluations and Emotions

Our approach focuses on the "cold" aspects of evaluation, if compared with others' (e.g. Mandler 1982), where the emotional components are particularly emphasized. Let us try to sketch our view of the complex relationship between evaluation and emotion.

5.1 *Evaluations do not necessarily imply emotions*

No doubt many evaluations show some emotional feature. For instance, if I believe a certain food, book, person to be good, I will be likely to feel attracted to it (him/her). But evaluations and emotions are not necessarily associated

with each other, because not *any* belief about the goodness or badness of something necessarily implies or induces an emotion or an attraction/rejection with regard to that "something". There also exist "cold" evaluations: if, for instance, I believe that John is a good typist, I will not necessarily feel attracted to him.

5.2 *Do emotions imply evaluations?*

This question is independent of the previous one, because, even if one admits that emotions imply evaluations, this would still leave room for "cold", i.e. non-emotional evaluations.

The answer to this question strictly depends on the level of analysis one may choose. To begin with, one may view attraction or rejection for some *x* as a (possible) *consequence* of an evaluation; so, in this case the emotion implies an evaluation in the sense we have just considered. On the other hand, however, one may view attraction or rejection *per se* as forms of evaluation of the attractive or repulsive object. In the latter case, we are dealing with a supposed identification: to say that an emotion implies an evaluation means to claim that the two actually coincide, which is still to be proved.

In fact, we view attraction and rejection as pre-cognitive implicit evaluations. We assume that a positive or negative emotional response can be associated with some stimulus. The automatic activation of this associated internal response (in Damasio's terms, a "somatic marker"; Damasio 1994) is the "appraisal" of the stimulus postulated by several theories of emotions (Arnold 1960; Frijda 1986; Lazarus, Averill, and Opton 1970; Smith and Ellsworth 1987). The associated negative or positive emotion makes the situation bad or good, unpleasant or pleasant, and we dislike or we like it. As a cognitive evaluation of *x* is likely to give rise to some goal (if *e* believes something *x* to be a means for *e*'s goal *p*, *e* comes to have the goal *q*, of acquiring and using the means *x*; see section 6.2), also the emotional appraisal of *x* gives rise to a goal: it activates a very general goal linked to the emotional reaction. This is the *conative* aspect of emotional appraisal. Positive appraisal activates an approach goal ("to be close to *x*; to have *x*"), while negative appraisal activates a generic avoidance goal ("not to be close to *x*; to avoid *x*").

We consider these sub-symbolic, implicit forms of "evaluation" as evolutionary *forerunners* of cognitive evaluations (and that is why we believe the answer to our question — do emotions imply evaluations — depends on the

level of analysis addressed). Indeed evaluation and emotional appraisal have much in common: their function. Evaluations favor the acquisition of adequate means for one's goals, and the avoidance of useless or dangerous means, and precisely the same function can be attributed to emotions.

More than that: emotions — though they have traditionally been attributed the negative role of clouding and altering rational thought — seem to *help* at least some kind of reasoning. In fact, they provide "nonconscious biases" that support processes of cognitive evaluation and reasoning (Bechara, Damasio, Tranel, and Damasio 1997), enabling for instance to choose an advantageous alternative before being able to explicitly evaluate it as advantageous.[2]

However, all this should not preclude to acknowledge the differences between emotional appraisal and cognitive evaluation, to address the latter in their own right, and try to establish their specific functions (see section 6.3). For instance, in some context emotional appraisal *by itself* might prove insufficient for assuring adaptive responses, in that, the more changeable and complex the world becomes (because of the increasing number of goals and contexts to deal with, and the complex relations among such goals and contexts), the more one needs analytical and flexible judgments about objects and events, rather than (or in addition to) more global and automatic reactions. In fact, evaluations allow to make subtle distinctions between similar (but not identical) goals and means, and to find out the right means for some new goal, never pursued in the past. Moreover, evaluations allow to reason about means and goals, and to *construct* and *transmit* theories for explaining or predicting the outcome of behavior.

Therefore, though emotional appraisal can be conceived of as an evolutionary *forerunner* of cognitive evaluation (as well as a valuable "support" for it), being an evolutionary "heir" does not imply maintaining the same nature as the forerunner; on the contrary, one might suppose that the same function has favored the development of different means, at different levels of complexity.

5.3 *Evaluations and subjectivity*

Evaluation is often viewed as an "affective", holistic, and subjective phenomenon. In social psychology, for instance, the notion of evaluation has been associated with such concepts as "reaction" and "interpretation" (e.g. Jones and Nisbett 1971). While perception and description are viewed as objective

phenomena, evaluation is considered similar to a personal and idiosyncratic reaction to a given object or event.

This can be traced back to at least three reasons: first, the kinship between evaluation and emotion, which, as just observed, has induced the undue identification of attraction or rejection with evaluation; second, the overlapping between evaluations and values (that we will discuss in section 8); third, the intuition that in evaluations the "self" plays a more significant role than in perceptions and descriptions (Mandler 1982).

In a sense, the latter assumption is right. The self is more involved in an evaluation than in a mere description of a given *x*, in that, while a description of *x* just includes (at least, at first sight) some beliefs about some characteristics of *x* itself, an evaluation implies another entity, that is, the evaluator, and his or her goals. It is with regard to such goals that *e* believes *x* to be good (or bad).

In fact, goals make our world "colored": as soon as some goal comes into play, the world it is no longer neutral, indifferent and objective; it becomes either positive (favorable), or negative (unfavorable), or indifferent (irrelevant to our goals). Generally this subjective polarization of the world is unduly and hastily traced back to emotions and to emotional appaisal. However it is primarily due to goals: it is only because organisms have goals that they have appraisals, "positive" and "negative" emotional reactions, and — at the cognitive explicit level — positive and negative evaluations of the world.

However, this does not imply that an evaluation should be subjective in the sense of idiosyncratic, emotional, and irrational. First of all, our notion of goal is quite distinct (and far) from that of wish, and from the domain of "wishful thinking" (e.g. Pears 1984). A goal is a mental representation with the role of a regulatory state, and *e* just evaluates *x* as a (good or bad) means for realizing that state. Moreover, as already observed, evaluations do not concern just the evaluator's goals: *e* may evaluate *x* with regard to someone else's goals, and even with regard to purely hypothetical goals. Finally, the objectivity and rationality of an evaluation are not hampered by the personal character of the goal under consideration: given a certain goal *p*, strange and idiosyncratic as it may be (say, to get ill), an evaluation of *x* (say, an excessive cold) as GOOD-FOR *p* is still a rational judgment.

Since the rational/irrational dimension is often at stake with regard to evaluations (and especially values, as we shall see), it may be worthwhile to discuss this aspect a little further. We assume evaluation is *typically rational*,

in the sense that its structure, that is, the means-end relation, is typically rational, if compared with other mental attitudes such as emotional appraisals or absolute judgments, where a conclusion is not derivable from its premises (and there is no way of verifying the judgment, or of arguing in favor or against it).

Of course, this does not imply that evaluations should necessarily be "right": assuming a relationship between a certain means and a goal does not entail that such a relationship should be real. However a *mistake* has little to do with irrationality. A wrong belief could be viewed as irrational if it is not grounded on any evidence or it is biased by some personal interest (Pears 1984). Let us consider the latter case, which is particularly relevant to our present concerns: if, for any reason, I *have the goal to believe* that certain nail scissors are good for cutting a piece of cloth, and I believe this *because of that goal*, then my evaluation will be irrational. Therefore, if it is true that some goal is implied in the production of an irrational evaluation, it is *the goal to believe* the evaluation as true, and not the goal mentioned within the evaluation itself (in our case, to cut the cloth). This equals to saying that evaluations are *not* intrinsically irrational. They can be irrational exactly as any other belief, including those of the descriptive kind, can be so, as far as they are affected by wishful thinking.

6. Evaluation and Action

Evaluations are a special kind of beliefs, characterized by a strict relationship with action, by virtue of their link with goals. But what do we mean by saying that evaluations are linked to goals?

6.1 *Evaluations imply goals*

Evaluations imply goals by definition, in that the latter are a necessary component of evaluations, namely, the second argument of the GOOD-FOR predicate. From a more "substantialist" perspective, evaluations imply goals in the sense that they *originate* from them: it is the existence of some goal p (either e's or someone else's) that justifies and motivates both the search for a means x to achieve it, and the belief that x is (not) GOOD-FOR p.

Potential evaluations

The crucial role of goals in evaluations is apparent if we consider that any belief, even the more neutral and descriptive one, is a *potential* evaluation, in that it turns into an evaluation as soon as some goal is considered, with regard to which the described object or event *ipso facto* acquires the status of a means. For instance the descriptive beliefs that "Eric is a German" or that "the rain waters the fields", turn into evaluations if considered with regard to the goals of, respectively, "letting Gertrud meet some German guy" and "having the fields watered".

Implicit versus explicit evaluations

Actual evaluations, that is, those evaluations which, not only potentially but in actual fact, contain the means-end relation, can present different degrees of explicitness.

An evaluation will be *explicit* if it directly expresses *e*'s belief about *x*'s power with regard to *p*, that is, if the GOOD-FOR predicate is directly mentioned in the belief. For instance a belief like "these scissors are (not) good (apt, suitable, etc.) for cutting a piece of cloth" is an explicit evaluation.

An evaluation will be *implicit* if the means-end relationship is given inferentially. Implicit evaluations directly mention those *qualities of x from which e has inferred the means-end relation*. For instance, "these scissors are sharp" is an implicit evaluation, as far as *e* is in fact implying that, because of this quality, they are good for some goal (otherwise, "these scissors are sharp" would just be a potential evaluation, like "these scissors are red").

In a sense, an implicit evaluation is richer than an explicit one, because it always implies the latter *plus* some information on *x*'s enabling qualities. The implicit nature of the evaluation lies in the fact that the relation between such qualities and *x*'s goodness/badness for *p* is not directly expressed in the proposition "these scissors are sharp".[3]

6.2 *Evaluations produce goals*

The relationship between evaluations and goals is even closer, because evaluations not only imply goals, but also can *generate* them. In fact, if *e* believes *x* is good for *p*, and *e* has that goal, *e* is also likely to want (possess, use) *x*. We propose a *rule of goal generation* which, using Cohen and Levesque's notion of *relativization* (Cohen and Levesque 1990a), might be expressed as follows:

(10) (BEL e (GOOD-FOR x p)) ⇒ (R-GOAL e (USE e x) (GOAL e p))

That is: if *e* believes something *x* to be a means for *e*'s goal *p*, *e* comes to have the goal (USE e x) of exploiting the means *x*. Such a goal is relativized (R-GOAL) to *p*, that is it will hold as long as (GOAL e p) holds.

In more synthetic terms, and going back to our definition (2) of an evaluation of a state or event *q* (rather than an entity *x*), (10) can be translated into:

(11) (BEL e (GOOD-FOR q p)) ⇒ (R-GOAL e q (GOAL e p))

However a number of specifications are needed. First of all, goal *q*, i.e. (USE e x), has not necessarily the status of an *intention* in *e*'s mind. For such a goal to become an intention, *e* should both choose to pursue *p* and prefer *x* to other possible alternative means for *p*. Moreover, a rule of this kind applies when *e* is the holder of goal *p*. When the holder of the goal is someone else, *e*'s evaluation will produce a *belief*, rather than a goal: the belief that agent *a* has goal *q* (provided *a* shares the same evaluation).

Finally, it should be stressed that (10) or (11) are a "weak" and purely descriptive form of the rule of goal generation: if *e* believes that this pair of scissors is good for cutting the cloth, and *e* wants to cut the cloth, then *e* will *happen* to have the goal of using the scissors. However, when something is thought to be good for something else, it not only happens to become a goal, but it also *should* become a goal. If *x* is a means for *p*, *x should* be acquired and used (provided that *x* is chosen among other possible alternative means, and *p* is chosen as a possible goal to pursue, that is, it is in the course of becoming an intention). Indeed, there exists a psychological difference between end-goals and means-goals, i.e. goals pursued in view of other goals. Means-goals have a constrictive touch; they are often tiring and boresome, but must be pursued if we want to achieve the superordinated goals. So, in a "strong" version of the rule of goal generation this "normative" character of the goals produced by evaluations should be expressed somehow. But, at the present level of analysis, we prefer to stay with the weaker and merely descriptive version of the rule.

6.3 *The function of evaluation: A bridge between beliefs and goals*

Evaluations, that is knowledge about "what is good for what", and "why", play a crucial role in all the cognitive activities which are based upon symbolic

and explicit representations, reasoning and deliberation. Let us briefly address problem solving and decision making. Further, we will consider the important role played by evaluations in emotions.

Evaluations in problem solving

Needless to say, knowledge and reasoning greatly improve the skills of a system capable to solve problems and build plans for achieving its goals. A system which is just endowed with goals and an action repertoire or fixed routines activated by its goals presents a number of obvious shortcomings, among which the inability to solve new problems, to find out new ways for satisfying old needs, and new destinations for old means and skills. On the contrary, a system endowed with general knowledge and reasoning skills can use such knowledge (in particular the knowledge about causal relationships between states of the world and events) to find out new solutions, because new goals and actions can be produced only from the interaction between action memory and knowledge.

However, such a production, even if intelligent and simulatively plausible in a number of cases, is still quite laborious. It would imply a continuous translation of knowledge into goals: as soon as a certain activated goal is seen as not achievable through the available actions, the system should search in its memory for possible causes of the desired effect, and translate causes into goals as many times as necessary, until it finds in the causal chain some effect it is able to establish through some of its actions.

As one can see, in such a context general knowledge is, so to say, neutral and static, without any direct relationship with the system's goals. The relationship between knowledge and goals has to be found and established every time a problem (i.e., a goal) arises.

The particular advantage offered by evaluative knowledge is precisely this relationship, in terms of beliefs about "what is good for what", derived from either one's experience about problems solved in the past, or one's reasoning and inferences, or others' communication. Evaluations make such a relationship explicit; they fill the gap between knowledge and goals, by reinterpreting the properties, qualities, and characteristics of objects and situations in terms of *means* for the system's (potential or actual) goals. The cognitive network ceases to be neutral and becomes "polarized" toward goals, that is ready for problem solving and decision making. It also becomes, so to say, more subjective, in that many descriptive traits of the world are attributed

some positive or negative value with regard to the subject's purposes and perspective.

Evaluations in preferences and choices

In a cognitive agent preferences can be internally represented both at the procedural and at the declarative (propositional) level.

Having a "procedural preference" means that, at a given level of their processing, a system's goals present different degrees or indexes of activation, priority, weight, value, importance (or whatever), that in fact create some rank order among them, which will be followed by some choice/selection procedure.

Having a "declarative preference" means that the system is endowed with an explicit belief like: "*x* is better than *y* (for *p*)". In particular, three types of beliefs are relevant for preferences: (a) simple evaluations, that is beliefs about how good/useful/apt/powerful are certain entities relative to a given goal ("*x* is very useful for *p*"; "*y* is quite insufficient for *p*"); (b) comparative evaluations like "*x* is better than *y* for *p*"; (c) reflective preference statements, of the kind "I prefer *x* to *y* (for *p*)". Generally, (b) are based on (a), and (c) are based on (b).

Both procedural and declarative preferences can coexist in a human mind (and would be of some use in artificial minds too), and each level of preference representation, though having its own mechanisms of reasoning, is translatable into the other. One can derive a "weight" from the declarative evaluations and their arguments, and *vice versa*, one can explicitly express (as beliefs) some priority of attractiveness, urgency, activation, or whatever.

However, being able to deliberate, that is, to choose an alternative on the grounds of explicit evaluations concerning the goodness of the various options, and being capable of reasoning aiming at supporting such judgments add further advantages to the mere fact of making choices. In these cases, in fact, the system can *justify* its choices, as well as *modify* the values at stake through reasoning. Moreover, it is liable to persuasion, that is, it can modify its preferences on the grounds of the evaluations conveyed by others.

It is worth emphasizing that evaluations can be independent of immediate actions. In order to retrieve information about the means for achieving a certain goal, it is not necessary for the goal to be actively pursued. The system might be hypothetically interested in that means-end relationship, and just want to know what could be done *in case p* were a goal. In addition, this knowledge can be subjected to the various kinds of elaboration (analogies, inferences, generalizations) that induce the multiplication of knowledge itself.

In other words, evaluations produce other evaluations, thus increasing the system's ability to solve problems and make decisions. Finally, a system of this sort, aimed not only at behaving but also at *knowing how to behave*, can transmit and receive evaluative knowledge, thus further expanding and modifying its equipment for problem solving and decision making.

Evaluations in emotions

Turning to less "rational" processes, let us now address again the evaluation-emotion relationship. Evaluations play a crucial role in the activation and constitution of emotional states. Most emotions, especially those typical of higher cognitive systems, are not just *triggered* by certain stimula (like an alarm reaction is triggered by a certain shape or color); they are rather *activated* by certain beliefs, which also become part of the emotions themselves. John cannot envy Tom unless he believes Tom has something he is unable to achieve; Tom cannot feel guilty unless he believes he is responsible for some wrong doing; and so on. As one can see, such beliefs are in fact of the evaluative kind. They imply the discrepancy or congruency between a certain state of the world and a certain goal. In fact, most theories of emotion emphasize the relationship between emotions and goals, and consider the former as signals for the system that its goals are either satisfied or threatened or frustrated (e.g. Arnold 1960; Frijda 1988; Higgins 1987; Ortony, Clore, and Collins 1988; Schwartz 1990). In cognitive terms, the expression of such a relationship is given by evaluations. Emotions are negative in that they signal the frustration of or threat to some goal; they are positive when they signal the (possible) achievement of some goal. A negative emotion implies (of course, as far as its cognitive components are considered) a negative evaluation; a positive emotion implies a positive evaluation. The cognitive study of emotion is in fact the study of the evaluations and evaluative processes underlying emotional responses.

7. Evaluation as a Condition for Social Action

Evaluation plays a crucial role also in social action. People have some "value" for other people, and the relationships among them are grounded on the "value" they attribute to each other. Sociality exists in that others are useful or dangerous with regard to our goals. People, like objects and events, are subject

to evaluations of their power, with regard to both their own goals and the goals of those who evaluate them.

The ability to deal with evaluations is a basic requirement for a cognitive agent that deals with other agents in a common world, and can achieve its goals through others' behaviors. The advantage of (cooperative) sociality is not only to speed up the search for solutions to common problems, or to find better solutions to them (Clearwater, Huberman, and Hogg 1991), but also to *multiply individual "powers"*: any agent, while remaining limited in its capabilities, skills and resources, finds the number of goals it can pursue and achieve increased by virtue of its use of others' skills and resources. But to deliberately use others' skills and resources one should be able to evaluate them.

Three individual evaluative abilities appear as particularly crucial in this context, in that they constitute the basic ways in which evaluations are conditions for social behavior:

- the ability to evaluate others' "powers" (competence, qualities, skills, resources);
- the ability to evaluate oneself in comparison with others;
- the ability to induce others' evaluations of oneself.

7.1 *The ability to evaluate others*

In a sense, any agent's limits of power, and its differences from others in the kind of power it is endowed with, turn into an advantage: although not omnipotent, the agent is allowed to overcome its computational, cognitive, and practical limits through "sociality". The agent is *dependent* on others for anything it cannot obtain by itself. This interdependence among agents is the basis for social behavior and in particular for cooperation and exchange (Castelfranchi, Miceli, and Cesta 1992). In a word, what any agent looks for in society is the *adoption of its own goals* by others.

But which "others" should one turn to? Who is the other who can help me, and for what goal? The knowledge for answering these questions is nothing but evaluations. When one assumes to be dependent on another, one is evaluating the other. In fact, one is assuming the other is endowed with the power one needs, and looks for the other's adoption (through help, cooperation, exchange) of its goals to use such power for those goals. It is evaluations which tell us who are the people we depend on, and those we should negotiate

with. Moreover these evaluations are of the *comparative* kind, because each agent is looking for the best value available in the market.

7.2 *The ability to evaluate oneself in comparison with others*

Evaluations are comparative also in the sense that each agent, while evaluating the others, involves itself in the evaluation, to assess the amount and kind of its own power in comparison with that of others. In a social world, in fact, "powerfulness" or "powerlessness" are comparative in nature: they imply having *more* or *less* power than a reference group or individual. Therefore each agent should be able to both evaluate oneself with respect to one's own goals, and to know how much "value" it has in comparison with others.

Moreover, an agent should know how others evaluate its power, thereby knowing its *rank* in the social hierarchy stemming from such (more or less) *shared comparative evaluations* (merits, reputation, prestige, and so on). One's *social power* depends on these comparative evaluations and on one's consequent position in the social hierarchy.

If agent A has the power to realize a certain goal and agent B depends on A for that goal, then A has some *power over* B, in that A is able to make B achieve or not achieve that goal. Also this is a kind of evaluation, which affects A's and B's respective positions in the social hierarchy, as well as their negotiations. By promising or threatening to use our "powers over others", we can obtain their adoption of our goals. In other words, our *power over* others is used as a basis for our *power of influencing* them (Castelfranchi 1990), that is the power to favor their doing or not doing what we want.

7.3 *The ability to induce others' evaluations of ourselves*

However, what most counts in social life is not the mere fact of being endowed with power (skills, competence, and resources) but the fact that *others believe* so. It greatly depends on others' evaluations of ourselves whether they will prefer to interact with us, accept our offers of cooperation or exchange, comply with our requests, and be influenced by our judgment and behavior. Agents look for each other because they need each other for satisfying their own goals (of any kind: from the more material and instrumental to the more affective and spiritual ones). If others *believe* I am not endowed with the power they need, if they believe I am not able to do what I promise or threaten

to do, they will not look for me, and will not try to interact and exchange with me. As a consequence, I will be socially powerless and outcast. Not surprisingly, competition among social agents often regards their respective images and consequent ranks in the hierarchy, rather than specific goods and resources.

This is the basic reason for the crucial role played by an agent's ability to present an *image* to the others such as to favor their adoption. It is worth emphasizing that such an image should not only imply others' positive evaluations of the agent's power in terms of competence. Others' positive evaluations should also regard the agent's reliability and trustworthiness. To appear as a good partner for exchanges, in fact, one should present oneself as both capable and willing to meet one's engagements and commitments, and to reciprocate others' help. In other words, one should not appear as a cheater or sponger.

As a consequence, a cognitive social agent will have a number of evaluative *goals*: the goal to evaluate both others and oneself, and compare such evaluations; the goal to know the others' evaluations of oneself; and the goal to present a good image, that is, to favor the others' positive evaluations of oneself. These goals guide a great deal of the agent's cognitive and social activity, and favor the collection and organization of a variety of evaluative beliefs. What is called the personal and social *identity* of an individual is mainly composed of complex structures of evaluations.

8. Values

In the social sciences, *value* seems to traditionally play the role of a passe-partout concept, accounting for a variety of phenomena, from the intrapsychic to the social domain. Time after time, it has come to overlap a number of concepts, such as: *valence* (Dalla Volta 1961; Pepper 1958; Perry 1954); *goal* (Koehler 1938; Pepper 1958), in particular of the general and long-term kind (von Cranach, Kalbermatten, Indermuhle, and Gugler 1982; Hollander 1967; Rokeach 1974); *need* (Maslow 1959); *standard* (Becker 1968; Parsons 1951; Williams 1951); and *norm* (Riecken and Homans 1954; Rokeach 1974; Williams 1964). Sometimes, values are also conceived as a kind of *belief*. The core of such views (often coexisting with one or more of the previous ones) is that values are "conceptions of the desirable" (Cooper

and McGaugh 1963; English and English 1959; Kluckhohn 1951; Krech, Crutchfield, and Ballachey 1968; Rokeach 1974; Williams 1964), where the *desirable* is often kept distinct from the *desired*; in other words, values would concern what *should* be desired, rather than what is actually desired.

By and large, we share the latter notion of values in terms of beliefs. But, before presenting our view in more detail, let us start from the meaning of *value* in everyday language, which in part accounts for the conceptual vagueness and overlapping we have just mentioned.

In everyday language, *value* has two distinct meanings, that are both related to the concept of evaluation. One meaning is "relative", that is the value of something *for something else*. This coincides with the result of an evaluation, or with the notion of *means* itself: if x is considered as a (good or bad) means for p, it acquires a (positive or negative) value with regard to p. By contrast, the other meaning of value is "absolute": for instance, honesty does not *have* a value in view of something else; it *is* a value. It is precisely this absolute meaning what we are interested in here.

When we say that x (say, honesty) is a (positive) value, we mean that it is *good*. But while in evaluating some entity x or state q as good, we are assuming it is good *for something else*, i.e. some goal p, here, on the contrary, we are considering x or q as *good per se*, or, more precisely, we are leaving unspecified what honesty is good for.

Therefore a value, in its absolute meaning, can be represented as an evaluation which leaves its second argument unspecified (Miceli and Castelfranchi 1989):

(12) (BEL e (GOOD-FOR x __))

Dewey (1939/1966: 5) identified two meanings of *valuing*, which imply the previous distinction between value and evaluation: *prizing*, in the sense of "holding precious", and *appraising*, in the sense of "assigning value to". Appraising is "an activity of rating, an act that involves comparison"; it concerns "a relational property of objects so that an intellectual aspect is uppermost of the same general sort that is found in '*estimate*' as distinguished from the personal-emotional word '*esteem*'". Ultimately, Dewey is suggesting a distinction between the objective and rational nature of some evaluations and the irrational and emotional quality of other "evaluations". We assume the latter are in fact values in their absolute meaning. Values are a special kind of evaluations where the "relational property of objects" is left in the mist of

potentials, and it is exactly for that reason that such evaluations can easily evade the control of rational (i.e., falsifiable) reasoning. Once the link between means and goals has been broken, it is impossible to prove that *x* is (not) good; if we do not know *why* (that is, *for what*) it is good, we cannot verify its goodness.

The means-ends dimension is also explicitly addressed by Dewey (1939/1966: 25), when he points out that "appraisal applies only to *means*, while prizing applies to things that are *ends*", that is, things that are good in and of themselves. We share this distinction, provided it is viewed in psychological (rather than "logical") terms. In fact, it should be stressed that the difference between values and evaluations is not a difference between "things good in themselves" and "things good for something else", respectively. The semantics of *good* always implies "for something". The difference between values and evaluations lies in the fact that *values leave unspecified* the "something" they are good for. Since the means-ends continuum is broken, values *appear* as absolute, with a number of important consequences (see section 8.2).

As indirect evidence of the fact that values result from a broken evaluation, consider that what is a value for some people can be viewed as an evaluation by other people, or even by the same people in different instances. In other words, what is given as an "end" in some propositions can be given as a "means" in others. For instance, one parent can consider studying (education, culture) as a value in itself, while another can view it just as a means for some goal (e.g. finding a job). A student, on the other hand, might consider studying as a means as well, but in view of some other goal (e.g. passing the exam; making dad happy; avoiding dad's punishment; and so on). With time, the student's perspective (that is, the reference goal) can change: he or she can move from one kind of evaluation to another, or even from an evaluation to a value, having studying as its object.

8.1 *Values, goals, and norms*

The notion of value we have suggested allows to distinguish it from related concepts which value is often made overlap with. Here we will consider goals and norms, whose "kinship" with values is particularly close.

Values and terminal goals

We have just said that values mention "means" without specifying the "ends"

of such means, and that is why they appear to mention "ends in themselves". In fact, values look very similar to a special kind of goals, i.e. *terminal* ones — those that, within the mind of the individual, are not represented as instrumental to other (higher order) goals, but as ends in themselves.

However, this should *not* make infer that values *coincide* with terminal goals. Goals are regulatory mental states, while values are just beliefs (in particular, evaluative beliefs of a special kind). In other words, values have a different status in the agent's mind: they are *judgments*, however general and absolute, about the goodness (preciousness, usefulness) of something. Consider a possible terminal goal such as "making friends", or "avoiding loneliness". As a terminal goal, "making friends" is a mental state that regulates my behavior (and makes me look for people, go to parties, etc.) as an end in itself (i.e. not in view of some other superordinate goal). As a value, "having friends" is, first of all, the object of a belief of the kind "having friends is good". Since this is what I believe, I will *also want* to have friends, but as a consequence of my belief. That is, *values are likely to generate goals*.

So, terminal goals do not coincide with values. They *may* be a *product* of values. However, not any terminal goal stems from a value. Going back to "making friends", I can have and pursue it as a terminal goal without necessarily believing it is a good thing to do. I might even harbor the value of solitude ("solitude is the price of greatness"), and regret going on with "making friends" and not being able to do without people.

Though not every terminal goal stems from a value, it is still true that every value is likely to give rise to some terminal goal. Such a strict relationship between values and goals is necessary according to our model of evaluation. In fact, according to our rule of goal generation (section 6.2), an evaluation of a given means (the use of an entity x or a certain state or event q) as good for a goal p "translates" that means q into a goal:

(11) (BEL e (GOOD-FOR q p)) ⇒ (R-GOAL e q (GOAL e p))

Here q is generated as *instrumental to* p; in fact, q is relativized to goal p: as long as e conceives p as a goal, e will have q as a means-goal. However, if we apply the rule of goal generation to a value, we will obtain a result of the following kind:

(13) (BEL e (GOOD-FOR q __)) ⇒ (R-GOAL e q (GOAL e __))

That is, since there is no specification of the goal in view of which q is conceived of as good, q is relativized to... nothing. This difference between

(11) and (13) implies an important consequence: while an evaluation produces an *instrumental* goal, a value produces a *terminal* goal. In fact, a goal that lacks its super-goal turns into an end in itself. So, since *q* is not relativized to any other goal, our rule could be rewritten in simpler terms:

(14) (BEL e (GOOD-FOR q __)) ⇒ (GOAL e q)

It is worth observing that there also exists a relationship between values and goals that goes in the opposite direction. Not only values generate (terminal) goals, but also vice versa, goals may "generate" values. Let us consider again the terminal goal of "making friends", and suppose that I have got it as such, i.e., it is *not* derived from a value of mine. I just feel this need for belongingness and friendship. However, at this point I can construct a corresponding value, and come to believe that "making friends" is a good thing to do. This is an ad hoc and post hoc generation of a value, that is meant to *justify* the existence of the terminal goal. Its psychological interest is quite apparent: in our search for supports and justifications for our needs and choices, we are likely to invert, so to say, the "natural" (or better, rational) order of the (mental) events; since what is good should be pursued, we want to believe that what we actually pursue is good. An old tradition of thought, starting from Spinoza, shares this view of values: we regard as good, and transform into values, what we like and desire.

Values and norms

What we have just said about the difference between values and goals also applies in the case of norms, when conceived of as *prescriptions*, that is a special kind of goals (socially shared, set on individuals in view of some common good, etc.). While norms, like goals, regulate behavior, values are a kind of evaluative beliefs. So, while "be reliable" would be a norm, "reliability is good" is a value.

When dealing with the rule of goal generation from evaluations (section 6.2), we have mentioned the "normative" facet of that rule: when something *q* is thought to be good for something else *p*, it not only happens to become a goal, but it also *should* become a goal. However, the normative character of a goal generated by an evaluation is, so to say, *conditional*: *if*, and *as long as*, *p is a goal*, *q* should be a goal. By contrast, values are more cogently normative, in that the goals they produce are unconditioned: since *q* is given as good *per se* (and not for some *p*), it *should* be a goal under any condition. So, the

terminal goal generated by a value in fact turns into a norm.

Such a close kinship between values and norms accounts for the common overlapping of the two concepts. And that is also why values are (rightly) seen as "conceptions of the desirable", as opposed to the actually "desired".

8.2 *General features of values*

Our analysis allows to account for a number of typical features of values, namely their being unfalsifiable, indefinite, normative, and terminal.

Values are unfalsifiable

If I do not know *why* something is good (that is, what goal it serves), I cannot *prove* it is so or *verify* if it is good or not. The goodness of something is meaningful only if it can be translated into its instrumentality for something else. Values, by definition, do not provide information about their own instrumentality.

Values are indefinite

Generally, objects, events, and behaviors present some *boundaries of application*, beyond which they are useless, or even detrimental: the use of an entity x or a state q can be good for p up to a certain point, or it can be good for p and at the same time bad for some other goal r. Undesired side-effects or conflicts among different coexisting goals generally establish the boundaries of application of entities, states, and events. However, the boundaries of values are hardly known. If I do not know what goals a certain x is good for, I will not know *how far* x is good, and where it can come into conflict with other goals and values.

The unfalsifiable and indefinite character of values places them in the realm of so-called "irrationality", which is crowded with unquestionable and fuzzy entities. In particular, the indefinite nature of values accounts for the fact that people are quite likely to harbor incompatible or conflicting values (e.g. "career" versus "family life"), without either realizing the existence of the conflict or being able to solve it (MacIntyre 1981). In fact, when a conflict is identified (because, for instance, one registers that two values cause conflicting side-effects), it can be very hard to solve it, because, not knowing which goals are served by the conflicting values, one cannot know what to prefer and choose.

Values are normative

As already shown, if *x* or *q* is good, it *should be wanted*, according to a rule of goal generation which, when applied to values, acquires a special normative character. It is worth observing that, if compared with actual norms, values present some advantages, in that they show a greater persuasive power. In fact, often it results more effective to convey norms through values, rather than to express them directly. Values appear more "acceptable" than norms for a number of reasons. First, values do not show the impositive character that is typical of prescriptions and commands. They do not say, "Do this!", but just express some judgment about the goodness or badness of something. Second, values *justify the prescriptions* they implicitly convey, that is, they provide some reason (i.e. *x*'s or *q*'s "goodness"), however absolute and vague, for complying with the implicit prescription. Third, values involve the conveyer: when conveying a value, one is also showing himself or herself to *share* that value, i.e the judgment of goodness or badness, and therefore also the prescription it implies.

Values are terminal

While the goals generated by evaluations are instrumental, in that they are relativized to the goals implied in the evaluations, the goals and norms generated by values are always terminal, i.e. ends in themselves. If I do not know which are the goals *x* or *q* is good for, I will look for it as an end in itself, not in view of something else. The normative and terminal nature of values accounts for their *dogmatic* character: values appear as unquestionable assumptions, that produce absolute imperatives. It is understandable that an individual's values are associated with or traced back to his or her relationships with "authority" and "significant others". Also the particular importance people attach to their values is quite understandable. Values can hardly be disregarded or neglected. The norms they convey cannot be violated, unless under pain of serious feelings of guilt.

9. Functions of Values

The possible function(s) of such dogmatic evaluations as values is not so apparent. The latter, unlike evaluations, do not suggest means for some goals, but just goals, without providing any explicit or grounded reason for pursuing

them. Ultimately, values are nothing but unproven judgments that claim to be taken for granted, which seems contrary to any wise and well-adjusted behavior. However, in spite of these apparent negative features, values can serve important functions, both social and cognitive.

9.1 *Social function of values*

Values are traditionally attributed the function of favoring and ensuring social stability and cohesion. In fact, the typical features of values we have just outlined account for such a function: on the one hand, values convey norms with the advantage of being more persuasive than norms as such; on the other hand, being unfalsifiable, values are also unquestionable. Therefore, they are beyond rational discussion and decision. One can just blindly and instinctively accept or reject them.

Once values have been introjected, it is very hard to get rid of them, as well as of the prescriptions they convey. This is what happens in the socialization period: the so-called introjection of "significant figures", including their values, occurs quite early in childhood, when the individual has little mastery of critical reasoning, and is very unlikely to question their validity.

9.2 *Cognitive function of values*

To account for the cognitive function of values, we have to go back to evaluations for a while. As we know, evaluations share some general properties of beliefs. For instance, as any other belief, an evaluation of an entity x or state q can be generalized to its class, and *vice versa*, an evaluation of x or q can be inherited from the evaluation of its class. As a number of different beliefs can regard the same entity or state, and give rise to some global concept, so a number of evaluations can regard the same entity or state, and give rise to a *global evaluation* of it. Let us consider this case in greater detail, starting again from our "scissors example".

A pair of scissors can be evaluated by e with respect to a variety of goals: to cut something, to make a hole in a wall, to kill somebody, and so on. Now, it can happen that e produces a global evaluation of the scissors on the grounds of its various specific evaluations. Such a "synthetic" evaluation, regarding a variety of goals, will necessarily leave unspecified its second argument:

(15) (BEL e (GOOD-FOR x p))
(BEL e (GOOD-FOR x q))
(BEL e (GOOD-FOR x r)) } ⇒ (BEL e (GOOD-FOR x __))

So, *e* produces a global evaluative belief about the goodness (or badness) of *x*, while leaving unspecified what *x* is good for. Of course, the more similar the starting specific evaluations — for instance, all of them are positive (or negative) and at the same degree — the more likely the synthesis to occur; the more different the starting evaluations — for instance, some of them are negative while others are positive — the less likely is the attempt to produce a synthesis in terms of a sort of algebraic sum of the various valences.

Such a synthetic and "broken" evaluation can still be supported by the analytic evaluations, meaning that *e* maintains in his or her memory the connections between the global evaluation and the specific ones. The global evaluation may play the role of a headline, a sort of generic summary of the various "GOOD-FORS". But it is also possible that *e* loses the trail of the starting evaluations, or even that *e* possesses the global evaluation as a given, having acquired it through social influencing and communication. In the latter cases, the global evaluation is a value.

We in fact assume values are either acquired as such (mostly, in the socialization period) or derived from a number of evaluations of the same entity or state of the world with regard to various different goals. This implies a generalization from the goodness of *x* for a variety of goals to its "absolute" goodness, that is, a broken evaluation which leaves its second argument unspecified. But what is the advantage offered by broken evaluations?

The multifunctionality of a given entity (i.e. its use for a variety of goals) not only justifies, but, so to say, requires the generalization of its various evaluations into a global one, for a number of reasons:

– a global evaluation is *economical*, in that it enables to relieve memory of a heavy work of retrieval of information. The greater the number of goals *x* is good for, the more likely *x* will be to serve the majority of goals in the majority of contexts. So, it is less expensive to believe that *x* is "good" *per se* and should be looked for in any circumstance, rather than trying to retrieve the specific evaluation (GOOD-FOR x w) which applies to a given goal and context;
– a global evaluation has an *open-ended potential*. The more *x* is multifunctional, the more unlimited its potential, meaning that *x* might have unexplored destinations, and serve new (not yet pursued) goals. In such

cases, it could be useful to believe that *x* is good *per se*, and apply it to the new context, rather than taking a "conservative" stance, and limiting oneself to considering the specific evaluations of *x* already obtained and verified;

– going back to socialization and learning, a simple and synthetic evaluation such as a value is *easier to acquire* than a variety of different specific evaluations of *x*. Here we are referring to the individual, rather than the collective, advantage offered by the social transmission of values. For a child it is easier to learn that *x* is good (or bad) and that it should be achieved or maintained (or avoided, refused, etc.), rather than to acquire and then retrieve the various evaluations about it, which are the reasons for such a generic "goodness".[4] The acquired value would not only reduce the load on the child's memory, but it would also prevent him or her from a difficult task of reasoning. In fact many evaluations are often just implicit, and should be inferred from other evaluations or from factual knowledge.

10. Concluding Remarks

In this work we have tried to outline a view of the cognitive role of evaluations and values, examining their features and relationships with the basic mental attitudes, i.e. beliefs and goals. We have emphasized the kinship between evaluations and values, on the one hand, and the relationship between any kind of evaluative belief and goals, on the other. Both evaluations and values are, to different degrees, *hybrid* mental objects, which mediate the relationship between factual beliefs and goals. The specific advantage offered by evaluations, if compared with other beliefs, is to make the cognitive network dynamic, by polarizing it toward problem solving and decision making, both in the nonsocial and in the social domain. Values show an even more hybrid and borderline nature (a sort of "advanced front" between evaluations and goals), which permits them to accomplish some special roles, both cognitive and social. Their being "broken", synthetic and global evaluations accounts for their cognitive functions, while their kinship with terminal goals and norms accounts for their social function.

Notes

1. A sort of precursor of evaluation by classification could be an evaluation "by analogy", where *x* is assimilated to another individual entity *x'* (which has been already tested with regard to *p*) rather than to a class *X* of entities. In other words, a generic similarity between the two entities makes *e* infer that *x* "will perform like *x'*" (which will allow *e* to do without an empirical test of *x*'s usefulness for *p*).

2. A number of studies conducted by Damasio and his collaborators (e.g. Bechara, Damasio, Damasio, and Anderson 1994) have pointed to the crucial role of emotion in cognitive evaluation and decision making. Their patients with lesions of the ventromedial prefrontal cortex show emotional blunting as well as difficulties in making decisions, especially in real-life contexts. If compared with normal subjects, they do not show stress reactions (as measured, for instance, by skin conductance response) when trying to make choices in uncertain and risky contexts (e.g. a gambling task). The interesting fact is that such emotional reactions, displayed by the normal subjects especially before making a wrong choice (i.e. a kind of choice previously associated with some punishment), help them to avoid it, and to opt for a less risky alternative. Such a choice is made *before* reasoning over the pertinent beliefs, including cognitive evaluations about the game, its options, and the possible strategies of decision making.

3. An implicit evaluation has indeed much in common with an evaluative standard. A standard is, so to say, the explicit translation of one (or more) implicit evaluations ("sharpness is an enabling quality of scissors with regard to the goal of cutting a piece of cloth"), together with the fact that the enabling quality is generalized to the class which *x* belongs to ("scissors" in general).

4. This stress on the functional aspects of values should by no means overshadow their (more obvious) negative implications, which are strictly related to their dogmatic and unquestionable character. Indeed, we believe that a closer relationship between values and the originating evaluations, and in general a critical attitude toward values, would be desirable, both at the social and at the individual level. More precisely, we believe that understanding or remembering where a value comes from, if pursued with respect to any specific instance, context or decision where that value applies, could be quite expensive, difficult, or even impossible. But it might be very useful for designing or modifying a social system, as well as a personal life-project. Retracing values' "steps" from possible evaluations could allow to identify the "reasons" for conflicts between values, or to suppress those values which are not (or no longer) justified by a real multifunctionality.

References

Abelson, R.P. 1979. Differences between Belief and Knowledge Systems. *Cognitive Science* 3: 355–366.

Agre, P.E. & D. Chapman. 1990. What Are Plans for? *Robotics and Autonomous Systems* 6: 17–34.

Anderson, J.R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Arnold, M.B. 1960. *Emotion and Personality*. New York: Columbia University Press.
Bechara, A., H. Damasio, D. Tranel, & A.R. Damasio. 1997. Deciding Advantageously before Knowing the Advantageous Strategy. *Science* 275: 1293–1295.
Bechara, A., A.R. Damasio, H. Damasio, & S.V. Anderson. 1994. Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex. *Cognition* 50: 7–15.
Becker, H. 1968. Value. In D.L. Sills (Ed.), *International Encyclopedia of the Social Sciences, Vol. 16*. New York: The Macmillian Company & The Free Press, 743–745.
Bratman, M.E. 1987. *Intentions, Plans, and Practical Reasons*. Cambridge, MA: Harvard University Press.
Bratman, M.E., D.J. Israel, & M.E. Pollack. 1988. Plans and Resource-Bounded Practical Reasoning. *Computational Intelligence* 4: 349–355.
Castelfranchi, C. 1990. Social Power: A missed point in DAI, MA and HCI. In Y. Demazeau & J.P. Mueller (Eds.), *Decentralized AI*. Amsterdam: Elsevier, 49–62.
Castelfranchi, C. 1995. Guarantees for Autonomy in Cognitive Agent Architecture. In M.J. Wooldridge & N.R. Jennings (Eds.), *Intelligent Agents*. Berlin: Springer-Verlag, 56–70.
Castelfranchi, C. 1996. Reasons: Belief support and goal dynamics. *Mathware & Soft Computing* 3: 233–247.
Castelfranchi, C., M. Miceli, & A. Cesta. 1992. Dependence Relations among Autonomous Agents. In E. Werner & Y. Demazeau (Eds.), *Decentralized AI-3*. Amsterdam: Elsevier, 49–62.
Clearwater, S.H., B.A. Huberman, & T. Hogg. 1991. Cooperative Solution of Constraint Satisfaction Problems" *Science* 254: 1181–1183.
Cohen, P.R. & H.J. Levesque. 1990a. Intention Is Choice with Commitment. *Artificial Intelligence* 42: 213–261.
Cohen, P.R. & H.J. Levesque. 1990b. Persistence, Intention and Commitment. In P.R. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in Communication*. Cambridge, Mass.: MIT Press, 33–71.
Conte, R. & C. Castelfranchi. 1995. *Cognitive and Social Action*. London: UCL Press.
Cooper, J. & J. McGaugh. 1963. *Integration Principles of Social Psychology*. New York: McGraw-Hill.
Cranach, M. von, V. Kalbermatten, K. Indermuhle, & B. Gugler. 1982. *Goal-Directed Action*. London: Academic Press.
Dalla Volta, A. 1961. *Dizionario di Psicologia*. Firenze: Editrice Universitaria.
Damasio, A.R. 1994. *Descartes' Error: Emotion, reason, and the human brain*. New York: Grosset/Putnam Book.
Davidson, P. 1996. *Autonomous Agents and the Concept of Concepts*. PhD Thesis. Lund University, Department of Computer Science, Sweden.
Dewey, J. 1939/1966. *Theory of Valuation*. Chicago: The University of Chicago Press.
Doyle, J. 1992. Rationality and Its Roles in Reasoning. *Computational Intelligence* 8: 376–409.
Dweck, C.S. 1991. Self-Theories and Goals: Their role in motivation, personality, and development. In R. Dienstbier (Ed.), *Nebraska Symposium on Motivation, Vol. 38. Perspectives on motivation*. Lincoln: University of Nebraska Press, 199–255.
Dweck, C.S. 1996. Implicit Theories as Organizers of Goals and Behavior." In P.M. Gollwitzer & J.A. Bargh (Eds.), *The Psychology of Action*. New York: Guilford Press,

69–90.
Dweck, C.S. & E.L. Leggett. 1988. A Social-Cognitive Approach to Personality and Motivation. *Psychological Review* 95: 256–273.
English, H. & A. English. 1959. *A Comprehensive Dictionary of Psychological and Psychoanalytical Terms*. London: Longmans Green.
Epstein, S. 1989. Values from the Perspective of Cognitive-Experiential Self-Theory. In N. Eisenberg, J. Reykowski, & E. Staub (Eds.), *Social and Moral Values: Individual and social perspectives*. Hillsdale, NJ: Erlbaum, 3–61.
Evans, J.St.B.T., S. E. Newstead, & R.M.J. Byrne. 1993. *Human Reasoning: The psychology of deduction*. Hillsdale, NJ: Erlbaum.
Frijda, N.H. 1986. *The Emotions*. New York: Cambridge University Press.
Frijda, N.H. 1988. The Laws of Emotion. *American Psychologist* 43: 349–358.
Galotti, K.M. 1989. Approaches to Studying Formal and Everyday Reasoning. *Psychological Bulletin* 105: 331–351.
Gollwitzer, P.M. 1996. The Volitional Benefits of Planning. In P.M. Gollwitzer and J.A. Bargh (Eds), *The Psychology of Action*. New York: The Guilford Press, 287–312.
Gollwitzer, P.M. & J.A. Bargh (Eds.). 1996. *The Psychology of Action*. New York: The Guilford Press.
Grosz, B.J. & C.L. Sidner. 1990. Plans for Discourse. In P.R. Cohen, J. Morgan, & M.E. Pollack (Eds.), *Intentions in Communication*. Cambridge, Mass.: MIT Press, 417–445.
Higgins, E.T. 1987. Self-Discrepancy: A theory relating self and affect. *Psychological Review* 94: 319–340.
Higgins, E.T. & R.M. Sorrentino (Eds.). 1990. *Handbook of Motivation and Cognition: Foundations of social behavior, Vol. 2*. New York: Guilford Press.
Hobbs, J.R. & R.C. Moore (Eds.). 1985. *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex.
Hollander, E.P. 1967. *Principles and Methods of Social Psychology*. Oxford: Oxford University Press.
Jones, E. & R. Nisbett. 1971. *The Actor and Observer. Divergent perceptions of the causes of behavior*. New York: General Learning Press.
Kluckhohn, C. 1951. *Values and Value Orientations in the Theory of Action*. Cambridge, Mass.: Cambridge University Press.
Koehler, W. 1938. *The Place of Value in a World of Facts*. New York: Liveright.
Krech, D., R.S. Crutchfield, & E.L. Ballachey. 1968. *Individual in Society: A textbook of social psychology*. New York: McGraw-Hill.
Kunda, Z. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108: 480–498.
Lazarus, R.S., J.R. Averill, & E.M.jr. Opton. 1970. Towards a Cognitive Theory of Emotion. In M.B. Arnold (Ed.), *Feelings and Emotions*. New York: Academic Press, 207–232.
Luck, M. & M. D'Inverno. 1995. A Formal Framework for Agency and Autonomy. *Proceedings of the First International Conference on Multi-Agent Systems*. Menlo Park, CA: AAAI Press, 254–260.
MacIntyre, A. 1981. *After Virtue: A study in moral theory*. Notre Dame, Indiana: University of Notre Dame Press.
Mandler, G. 1982. The Structure of Value: Accounting for taste. In M.S. Clark & S.T. Fiske

(Eds.), *Affect and Cognition*. Hillsdale, NJ: Erlbaum, 3–36.
Maslow, A.H. (Ed.) 1959. *New Knowledge in Human Values*. New York: Harper & Bros.
McDermott, D. 1987. A Critique of Pure Reason. *Computational Intelligence* 3: 151–160.
Medin, D.L. 1989. Concepts and Conceptual Structure. *American Psychologist* 44: 1469–1481.
Miceli, M. 1992. How to Make Someone Feel Guilty. Strategies of guilt inducement and their goals. *Journal for the Theory of Social Behaviour* 22: 81–104.
Miceli, M. & C. Castelfranchi. 1989. A Cognitive Approach to Values. *Journal for the Theory of Social Behaviour* 19: 169–193.
Miceli, M. & C. Castelfranchi. 1992. *La Cognizione del Valore*. Milano: Franco Angeli.
Miller, G.A., E. Galanter, & K.H. Pribram. 1960. *Plans and the Structure of Behavior*. New York: Holt.
Murphy, G.L. & D.L. Medin. 1985. The Role of Theories in Conceptual Coherence. *Psychological Review* 92: 289–316.
Oaksford, M. & N. Chater. 1991. Against Logicist Cognitive Science. *Mind & Language* 6: 1–38.
Ortony, A., G.L. Clore, & A. Collins. 1988. *The Cognitive Structure of Emotions*. New York: Cambridge University Press.
Parsons, T. 1951. *The Social System*. Glencoe: The Free Press.
Pears, D. 1984. *Motivated Irrationality*. Oxford: Oxford University Press.
Pepper, S.C. 1958. *The Sources of Value*. Berkeley: University of California Press.
Perry, R.B. 1954. *Realms of Value: A critique of human civilization*. Cambridge, Mass.: Harvard University Press.
Pylyshyn, Z.W. 1984. *Computation and Cognition*. Cambridge, Mass: MIT Press.
Rao, A.S. & M.P. Georgeff. 1991. Modelling Rational Agents within a BDI-Architecture. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*. San Mateo, CA: Morgan Kaufmann Publishers, 473–484.
Riecken, H.W. & G.C. Homans. 1954. Psychological Aspects of Social Structure. In G. Lindzey (Ed.), *Handbook of Social Psychology, Vol. 2*. Cambridge, Mass.: Cambridge University Press, 786–832.
Rokeach, M. 1974. *The Nature of Human Values*. New York: The Free Press.
Rosenblueth, A., N. Wiener, & J. Bigelow. 1968. Behavior, Purpose, and Teleology. In W. Buckley (Ed.), *Modern Systems Research for the Behavioral Scientist*. Chicago: Aldine, 368–372.
Ross, M. 1989. Relation of Implicit Theories to the Construction of Personal Histories. *Psychological Review* 96: 341–357.
Schwartz, N. 1990. Feelings as Information: Informational and motivational functions of affective states. In E.T. Higgins & R.M. Sorrentino (Eds.), *Handbook of Motivation and Cognition: Foundations of social behavior, Vol. 2*. New York: Guilford Press, 527–561.
Simon, H.A. 1979. *Models of Thought*. New Haven: Yale University Press.
Sloman, A. 1987. Motives, Mechanisms, and Emotions. *Cognition and Emotion* 1: 217–233.
Smith, C.A. & P.C. Ellsworth. 1987. Patterns of Appraisal and Emotion Related to Taking an Exam. *Journal of Personality and Social Psychology* 52: 475–488.

Sorrentino, R.M. & E.T. Higgins (Eds). 1986. *Handbook of Motivation and Cognition: Foundations of social behavior, Vol. 1*. New York: Guilford Press.

Steels, L. 1987. The Deepening of Expert Systems. *AI Communications* 1: 9–17.

Tverski, A. & D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and biases. *Science* 185: 423–469.

Wason, P.C. & P.N. Johnson-Laird. 1972. *Psychology of Reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

Williams, R.M.Jr. 1951. *American Society: A sociological interpretation*. New York: Knopf.

Williams, R.M.Jr. 1964. The Concept of Values. In J. Gould & W.L. Kolb (Eds.), *A Dictionary of the Social Sciences*. Glencoe: The Free Press, 283–287.

Wittenbrink, B., P.L. Gist, & J.L. Hilton. 1993. *The Perceiver as Alchemist: Conceptualizing stereotypes as theories*. Paper presented at the Fifth Annual Convention of the American Psychological Society, Chicago.

CHAPTER 10

Ontogeny of the Social Self

Towards a Formal Computational Theory

Eric Werner *

1. Introduction

How does an agent become social? How does a society continue to regenerate itself? We take a perspective that tries to balance a psychological account of the development of the self of an agent with a purely external, sociological account. For the purposes of social theory we need to have a theory of the agent that makes social action possible. For that to be the case we need to have agents that can participate in cooperative, role governed, organizational activity. At the same time, we want to keep our ontological presuppositions to a minimum. Thus, we do not want to assume positivistic, external laws that guide the agents as, for example, physical objects are guided by Newtonian laws. Instead, agents are guided by their intentional states including their reactive strategies. These intentional states are constrained by norms. Norms, however, must be internalized to have an effect on the agent's actions. In this essay, we present a theory of internalization of social abilities. This is the basis of a theory of socialization of an agent. The socialization of an agent is the basis for the transfer of culture of one generation to the next generation of agents. It is this socialization that sustains the very possibility of social activity.

---

* © Werner 1994, 1995, 1996, 1997, 1998, 1999. All rights reserved.

A theory of the development of the social psyche of an agent is presented. The purpose is the elucidation of the problem of how an agent is integrated into a social system. The social self is that part of the self that gives the agent a social competence. As part of the quest of understanding how an agent's social competence develops, we give a formal description of the process of internalization of social roles. This is made possible by utilization of a formal theory of an agent's intentions, social roles, as well as a formal theory of an agent's abilities in a context of a social world. See (Werner 1988 1989, 1994) for the formal theories of agent intentional states, multi-agent communication and agent coordination that provide the conceptual and formal foundations for this paper. This paper is also motivated and supported by psychological theories of the development of the self of the child in the maternal context. In characterizing this process we hope to further our understanding of how agents can learn and develop social abilities dynamically in a social world.

This essay has had a life as underground scientific literature for the past several years. I keep getting requests for this essay even though officially it has never been published and the original copies were marked "Not to be distributed!". It also has evoked diverse extreme positive and negative emotional reactions in readers. Some are upset at attempting to formalize aspects of the self. Others want more formalism. Still others raised on a diet of traditional mathematics want differential equations. My approach follows the tradition of logical formalism, semantics and pragmatics. Sociologists want more of their version of sociology. Psychologists want more of their own favorite brand of psychology. Still most, including computer scientists, tend to like the interdisciplinary knitting together of such divergent concepts and theories. While I did not want to publish it in its present form, since it is speculative and preliminary, I thought it would be more fair and fun to make it available for the broader scientific community. The essay may be read by ignoring the formal parts, although the comprehension may be more limited as a result.

2. Internal Representation of the Social System

Agents acting in the context of a social system have a partial representation of the intentions of the social system. This partial representation of the social system need not be a symbolic representation. By representation I mean that

the agent has information about the social system that the agent can utilize in its decision making. Thus, representation is neutral with respect to the kind of realization that this information has in the agent.

3. The Development of the Self

The self of the agent is built up in the social matrix, e.g., the mother-child relationship. At the earliest level the child's cries and behavior are interpreted by the mother, and she responds on the basis of that interpretation with actions that bring the child back in a state of balance. The child, in turn, builds up a representation of the mother as agent that responds to his gestures. This connection of gesture and strategic action on the part of the mother leads to the association of a gesture with its strategic meaning. The strategic meaning of a gesture is the strategy of the other agent in response to the gesture in various contexts. This strategic meaning is built up incrementally by the child. Behind the scenes is a social competence or social capacity that enables the child to represent social agents and their relations. Note, we will use the term 'mother' for a helping agent and 'child' for the agent who is being helped.

So far we have not defined the self. The self of the agent consists of the information state, strategic state and evaluative state with regard to the agent as center of volition and action. The representation of self in part determines the agent's strategic decisions and actions. In the beginning the agent may not distinguish between himself and the helping agent. The child may not realize that the mother is a distinct agent and not just an extension of himself.

Thus, for example, initially the child may view the mother as enslaved to his wishes and controlled through gestures, much as his hands are enslaved to his volitional control. To the extent that the mother is good at guessing the wishes of the child, there will be a fusion of self and other. There will however be a representation of the physical mother, in the space-time possibility space, that is different from the physical self of the child. The child can observe that the mother moves even when he has no volition or made any gestures in that regard. But, much like a pet dog, as soon as the child has an urge the mother will be there.

This fundamental relationship of need and response to need, builds up trust and confidence which is a representation of the other as being capable and cooperative when she is needed.

4. The Self and Internalization

As the child gains more inherent abilities the adequate mother will realize this and relax her vigilance. There will be small frustrations for the child where the mother fails to guess the meaning of a gesture, or where she does not want to or cannot fulfill the strategic content because of other goals. To the extent that the frustrations are small and fit within the growing capacities of the child, the child utilizes his representation of the mother's strategic response to his gesture by identifying with the mother and taking on her role. This process is called *internalization*. In internalization the child utilizes the strategic representation of the other to guide his own actions in fulfilling his own goals. Thus, an expectation is converted to a control strategy.

In the process of internalization whole blocks of personality are integrated into the self. These blocks consist of state information (beliefs), strategic information (intentions, goals) and evaluative information (wants, wishes, emotional reactions). What happens is an internal change. The representation of the other becomes part of the representation of the self and in doing so changes the control state of the agent. The self of the agent actual changes by this transformation.

Note, it is an open question if this internalization results in a doubling of the representation of the other, one for the other and one for the self, or if this information is merely accessed to control the self. The access may occur by new connections to the representation of the other that permit this representation to be utilized in the service of the self. We will distinguish *copy internalization* where a copy of a strategic, intentional structure or capacity of the other is copied to become part of the self, from *reference internalization* where no copy is made but rather the same strategic, intentional capacity of the other is referenced or pointed to, in order to be accessible as a possible control strategy in the service of the self. There are fundamentally different properties holding of these two types of internalization processes. We will discuss this in more detail when we look at the dual and contrary process of externalization.

5. A Formal Theory of the Process of Internalization

5.1 *The representational state of the agent*

If we describe this more formally the cognitive state R_C of the child-agent C, then C starts in a state of total uncertainty of state σ, plans π and values ψ. Thus, $R_C = < I_C; S_C; V_C >$ where I_C is the information the agent C has about the state of the world, S_C is the information about the agent's plans and intentions (including the uncertainty about the intentions of others), and V_C is the information about the agent's values. As the child interacts with the natural and social world, the perceptions of the child operate on the information state I_C. Indeed, we can associate with a perception p an operator θ_p that updates and transforms the information state I_C to a new state $\theta_p(I_C) = I'_C$. This new information state I'_C reflects the new knowledge gained by the child through the perception p.

5.2 *The construction of the representation of the other*

As the child-agent C gains information about the strategic responses of the mother-agent M to the gestures g of the child, the child builds up his strategic representation S_C^M of the mother. In other words, the child's perception p_{r_M} of the response r_M of the mother M to the child's gesture g_C modifies the child's representation S_C^M of the strategies of the mother as follows: $p_{r_M}\left(S_C^M\right) = S_C^{M'}$ where $S_C^{M'}$ is the new strategic information state that the child has about the mother's intentional state after the child's perception $p_{r(M)}$ of the mother's response r_M. Again, perceptions are treated as operators on the representational state R_C of the agent C. To be precise, we should really separate the perception p from the agent's interpretation of the perception. The *interpretation* is the operator associated with the perception. For simplicity we have identified the perception operator with the interpreting operator.

The strategy that the child builds up to represent the intentional state of the mother is a gesture-action strategy. It is, thus, a nonhomogeneous strategy that fuses language like gestures g_C of the agent C with action responses r_M of the agent M. These might be programmed as rules of the form: if g_C then r_M.

5.3 *The process of internalization*

Let us say we have the situation where the child, through long interaction with the mother, has built up a reliable, predictive strategic representation S_C^M of the mother's gestural response strategy.

The process of internalization can now be described formally as follows: The child C utilizes the strategic representation S_C^M it has of the mother M by identifying with his mother. In this process, parts of the representation of the mother S_C^M is transformed to be part of the child's representation of himself S_C^C. In other words, the process of internalization transforms the agent $C's$ control state S_C to S'_C. Here, $S'_C = S_C \otimes S_C^{[C/M]}$ and $S_C^{C/M}$ is the result of the substitution of the child C for the role of the mother M in S_C^M. This substitution corresponds to the child's identification with the mother. Here $\otimes$ is a general updating and fusion procedure. The formalization and implementation of this fusion operator $\otimes$ will vary with how we define strategic states. For now we leave this open. Details can be found in (Werner 1988, 1989, 1994). Thus, in the process of identification and internalization the child's intentional, strategic control structure S_C transforms into S'_C, i.e., if C internalizes S_C^M then $S_C \rightarrow S'_C = S_C \otimes S_C^{[C/M]}$. This same process of identification and internalization also can be applied to state information I_C^M and evaluative information V_C^M that the agent C has of M.

In general, if $R_C^M = < I_C^M, S_C^M, V_C^M >$ and C internalizes this state, strategic and evaluative information, then the agent's given representational state is changed to:

$$R'_C = < I_C \otimes I_C^{[C/M]}, S_C \otimes S_C^{[C/M]}, V_C \otimes V_C^{[C/M]} >$$

Looked at more abstractly, the internalization process can be viewed as an operator on the representational state of the agent. If we denote the *internalization operator* by ι then ι: $R_C \xrightarrow{\iota} R'_C$.

Looked at in more detail, the internalization *(shift internalization)* process $S_C \rightarrow S'_C = S_C \otimes S_C^{[C/M]}$ consists of two parts:

1. A gain of function in the child: $S_C = S_C \otimes S_C^{\text{role}_M}$

2. A loss of function in the mother: $S_C^M = S_C^M - S_C^{\text{role}_M}$

Here the loss of function of the mother relates to the mother's refusal and the concomitant loss of expectations on the part of the child toward the mother as helping agent, to respond to the gestures of the child. In fact the mother is still capable of performing the former service for the child. The refusal, however, is tantamount to a loss of function. The other is no longer represented as willing or needed to do the service. There is also a positive joy pride in the child in self volition, control and in newly acquired capabilities. Thus, there is active copying and encouragement and not just being forced into autonomy.

In this process of internalization there is a shift from a representation of the other's intention to utilizing that strategic plan information to form the agent's own intentions and goals. Therefore, after the internalization process the agent is guided by those new, previously external intentions. His control state S_C has been altered. His behavior, in so far as it is affected by the new internalized intentions, will be changed.

The internalization operator ι may be realized by many small, incremental steps or sub operators that gradually build up the new intentional structure of the agent. On the other hand, larger steps would explain discontinuous maturational steps including rapid increases in social capacities such occur in certain periods of a person's life. This would happen if the agent C has a good representation S_C^M and under certain conditions takes in this complex representation in large chunks to form the more complex intentional state S_C.

Note, that internalization is not limited to the mother or father, but may include any significant other in the child's social life. Thus, one may commonly observe a child gaining the attributes of one of his close friends. Indeed, parents and cultures restrict the peer group of children to prevent unwanted internalizations. Furthermore, in humans such internalization processes need not be restricted to childhood, but may occur in any phase of life.

5.4 *Internalization and the boundaries of self and other*

In the internalization process there is a shift of the boundary of the self. The self expands and incorporates the intentions (beliefs and values) of the other. The self also changes in the sense of shifting and not just growing. Infantile strategies, intentions and goals are given up or modified in favor of new internalized forms of behavior. There are also new capacities which allow the realization of a strategy and, therefore, of new intentional states. The self

expands its strategic capacity. The self can do more than before. Since the semantics of the logic of ability and 'can' is based on the existence of strategies (Werner 1991), internalization leads to an increase in abilities. Indeed, the very capacity to assume a role requires the internalization of the strategies associated with the role. To participate in the coordinated interaction in a group the capacity to assume such a role is essential. There the individual temporarily takes on an intentional substate to fulfill the role. Role playing is a temporary acceptance of a different self.

Is the self of the child just a sub intentional state of the intentional capacity of the adult? Is there a mapping from the intentional state of the child to a intentional substate of the internalized parent? The child utilizes the internalized strategies to further its own goals. In that sense the child has a sub self of its parents that is in the service of its own goals. These goals may themselves be the result of internalization. But the key difference is that the parameters, the indices refer to the child and not to the parents. Thus, even if the child is a straight copy of a portion of its parents intentional state, the variable 'I' refers to the child.

The different goals based on different wants and needs make different use of the same plans. The child uses his learned ability to walk, to go play or to find the teddy-bear. The child's local evaluations (wants, emotions) insure the preservation of self, and not just the structure of the intentions, since that structure, as we have seen, may be indistinguishable as a schema from that of the parents. As the result of internalization the child may eat exactly like the parents, but it is feeding itself and not the parents.

There is the further effect that the child has many agents as models in the internalization process. Thus, the child will have an intentional structure that is a mixture of the parents and others as well.

6. A Theory of Externalization

Externalization can be viewed as a complementary psychodynamic process to internalization. The externalization process takes a given representational strategic state and projects it onto the other. Yet, it is more than just projection since the actual control strategies of the self are modified. In the case of the child, the child goes into an earlier mode of behavior by externalizing functions it has previously mastered and internalized. The new state S'_C results

from S_C by subtracting a substrategic capacity of the child C and adding it as part of the strategic capacity of the mother M.

More formally, externalization $S_C \rightarrow S'_C = S_C - S_C^{[M/C]}$ consists of two parts:

1. A loss of function in the child: $S_C = S_C - S_C^{\text{role}_C}$

2. A gain of function in the mother: $S_C^M = S_C^M \otimes S_C^{\text{role}_M}$

6.1 *Types of externalization*

Looking at the phenomenon in more detail, in *externalization* the self projects parts of the self onto an external, actually existing object. There is a complication here because the *actual object* has a representation in the self. Let us call the representation the self has of the actual object the *external object.*

In the process of externalization, a transformation occurs that is the reverse of internalization. With regard to strategic capabilities, externalization may result in a shifting of some of the strategic content residing in the self to the external object. In fact we might distinguish three kinds of externalization:

1. *Copy externalization* where a copy of certain strategic representations is made to form parts of the external object. A transformation of the intentional state of the external object would be local to the external object.
2. *Reference externalization* where a reference (or pointer) is made from the external object to certain strategic representations, some of which may reside in the self. A transformation of the external object's intentional state need no longer be local; a change in the external object's strategies then would result in a transformation of the intentional state of the self to the extent that the referenced strategies reside in or are also referenced by the self.
3. *Move or Shift externalization* This can be realized as a combination of external object reference externalization and a dereferencing of the self from those referenced strategic representations. A transformation of the external objects referenced strategies would then no longer result in changes in the intentional behavior of the self. Hence, such transformations of the external object would be local to that object.

Note, though, that the possibility still exists that there are nonlocal effects on other external objects if these are not properly dereferenced or copied.

Move externalization can also be realized or implemented as a copy externalization followed by a delete of the copied parts in the self.

We could have a *library of strategies* or more generally a strategy base with a strategy generator using some kind of combinational logic/semantics. These components can be copied or generated locally for an object (self, other, external) where transformations of the components then have local effects or they can be created by generation for the library but then referenced by objects. In this case any transformation in a component will have nonlocal effects on all the objects that reference the component.

These are normal everyday phenomena and occur when taking on roles or abdicating from such. Role formation has both aspects of externalization and internalization. We can also observe such phenomena in persons that have boundaries that are partially fused with the other. This can occur as the result of internalization or by reference externalization.

7. Dynamic Boundaries of Self

7.1 *Self and other*

The differentiation between self and other has at least three fundamentally different aspects. First, there is the differentiation in space-time of the body of the self from the body of the other. Second, there is the differentiation between the identity of the self as agent or center of motivation and action from that of the other as independent agent. Third, there is the differentiation of the values of the self as agent and the evaluations that the other makes. This is related to the differentiation between the emotions of the self as agent from the emotions of the other. For example, empathy for the other is different from the lack of differentiation between the emotions resulting from a state the self is in versus the emotions resulting from a state that the other is in. Thus, if some hurtful situation occurs to the other, an emotionally undifferentiated agent, one who is emotionally merged with the other, will respond to the hurtful event as if it occurred to himself.

7.2 *Intrapsychic life of partners*

In relation to cooperation in dyads or groups, one can observe a continual

dynamic intrapsychic process of internalization and externalization. Indeed the cooperative act can be viewed from this perspective. The self in relying on the other to perform an action that it could also do itself, gradually externalizes that action to the domain of the other. There are many symbolic acts of cooperation between male and female that are strictly speaking not necessary in the sense that the actions can be performed by the individual parties themselves. In some relationships one can observe what some might term regression where there is a gradual diminution of function in one or both of the partners with the complementary role being taken over by the other. The other in this case internalizes new intentions while the self externalizes these strategies with a resulting loss in capabilities. If this is done in a symmetric, complementary way we observe a unity or partnership. If it is too one-sided we observe a strong dependency relationship. The point is that the self of each of the actual persons has lost some of its capacities in the service of the cooperative relationship.

A good example of this is the mutual dependencies and loss of capacities that occur in many marriages. The husband may loose many of his capacities to care for himself, to the extent that many women feel they have another child on their hands. So too the wife may loose so many of her abilities of self volition and direction that the husband may feel he has a child on his hands. Indeed, one often observes each of the partners parenting the other. To the extent that these capacities are actually move externalized, we observe mini regressions in the self in specific marital contexts.

7.3 *Internalization and loss*

In the case the actual other is lost through death or through unavailability (abandonment, long absence), the externalization process is (gradually) reversed with the lost aspects of the self being reinternalized. The nature of the internalization determines what effect it has on the external other. Internalization as the inverse of externalization can be of three types as well *copy-internalization, reference-internalization, and move-internalization*. Perhaps it is the maturity of the self, its nature or internalization/externalization method that determines the effects on the external object and the self and, therefore, on the nature of the resulting relationship between self and other.

If the self internalizes by movement, the external object is gradually lost. Persons suffering from this method of internalization fear abandonment since

it means the actual loss of the other as external object. The return of the other can be experienced by the self as the return of a stranger since the external object has been destroyed in the mean time.

If the self internalizes by reference, then the external object is not lost. However, the object is gradually changed, since the transformation of those internalized intentions is not local to the self. In other words, the transformations have a side effect on the representation of the other. The return of the actual object may result in a surprise in the differences of the other. Not just because the actual other may have changed, but also, because the external object (the representation of the other) has been transformed.

7.4 *Regression and loss of the self*

Notice, that externalization by movement would result in a gradual loss in the capacities the self. If this externalization is sudden and massive then we might speak of a loss of the self. This sort of process appears to underlie the phenomenon of regression. If the movement is irreversible we have a permanent loss of capacities and if the movement is in addition massive then we would observe a permanent loss of the mature self with severe and lasting regression.

Indeed if the process works by replacement where higher level strategies replace lower level strategies then loss of higher strategies by movement would result in the re-emergence of lower level strategies, i.e., regressive behaviors.

It is a well known phenomenon in children and in critical periods of life in adults that regression occurs regularly in the process of growth and reorganization. In this process strategies may be externalized, leading to regression to earlier strategies. Then a period of consolidation follows where new strategies are built up combinationally from more primitive ones. These are internalized and replace the more primitive, regressive strategies. This pattern would explain the waves of maturity and immaturity in the developmental process of children. Furthermore, it would explain the ebb and flow of emotional states in adults when faced with critical situations they cannot at first master.

8. Communication and the Self

8.1 *Communication and the formation of self*

One agent may influence another through communication. We have described in (Werner 1988a,b, 1989) how when a directive $\alpha!$ is understood it acts on the hypothetical intentions of the listener by a pragmatic operator $P(\alpha!)$. If the directive is accepted then the pragmatic operator acts on the actual intentional state S of the hearer transforming it to $P(\alpha!)(S)$. Thus communication can change the intentional state of an agent. Adults use directives quite frequently to change the intentional state of the child. Thus, for example, "Please pick up your shirt.", "Don't do that!" or "No!" or "Come here!" affect the child. The child will often respond ambivalently if the directive is at odds with its interests. Thus one must distinguish the pragmatic meaning of a communication from the actual, perlocutionary effect.

8.1.1 *Internalization of linguistic strategies*

However, even the directives themselves, or, more accurately, linguistic strategies that contain directives, e.g., commands and requests, are internalized by the child. So that one can observe a child giving directives to itself when he is in a similar situation. Thus linguistic intentional structures are also the objects of internalization. The internalization of communicative protocols and social interaction styles leads to the possibility of cooperative social interaction of the child in the social world. The protocols are local roles in an interaction where the strategic coupling of the two roles leads to a social, coordinated interaction. This coordinated interaction is the basis of social cooperation which, in turn, leads to the achievement of social events.

8.1.2 *A problem of other to self role transfer*

There is a theoretical problem: How can the child learn a role that is a dual or complimentary role when the parent only presents one side of the interaction protocol? If the process of learning is through internalization of observed and represented intentions, if the complimentary intentional structure is never observed, how can it be internalized? First, a parent often plays both sides of the role. Second, the child observes both roles being acted out by two other agents (parent-parent, parent-adult, parent-other-child). Third, the role may stimulate a more basic already known strategic reaction. These reactions form

the basis of the complimentary social role. Fourth, a role may evoke evaluative responses that lead to typical reactions that are then codified by repetition. Fifth, there is the process of parental mirroring. The parent mirrors the facial gestures of the child. The child then internalizes the reaction to its gestures. Once, the child has internalized the reactive gestures, it can then react to the adults gestures using the mirroring strategy. That then builds the basis of further more complex role interactions. Finally, there may be a genetic endowment of social interaction protocols that are used to start the process of social protocol internalization. The capacity of the child to respond to facial gestures at all, may already utilize some complementary, basic social roles of interaction.

8.1.3 *Learning a social language is a social process*

From the formal perspective there is then a gradual transference of the intentional structure of the social world to the social mind of the child. Linguistic interaction is just one kind of intentional, strategic, social interaction. The learning of social roles may be as crucial to proper linguistic interaction as the learning of the vocabulary. For what is learned is not just the grammatical, syntactic structure of sentences, but when to assert sentences in what social moment and in response to what gesture. Thus whole protocols or linguistic roles are learned and not just a generative capacity to spew out grammatical sentences. The Gricean maxims (Grice 1975) are a part of the cooperation strategies in linguistic communication. For they are principles that describe cooperative linguistic behavior.

Learning to cooperate is, therefore, a process of learning social roles that are complementary. This process is often mediated by prior social roles. There is then an ontogeny of social interaction skills that leads to more and more complex intentional structures.

8.2 *Communication with the other in the service of the self*

8.2.1 *Action and communication: Vygotsky*

An agent in a social world may utilize language to get another agent to help him in his task. To do this the agent utilizes a representation of his own abilities and those of the other. Cooperative action then builds a synthetic whole consisting of self-action, communication and other-action. A good example is provided by Vygotsky (Vygotsky 1987: p. 27), where a child

attempts to solve a problem by communicating with himself (egocentric speech for planning and observation), performing actions (plan execution), and communicating with the observing adult (social speech) to get assistance when he is unable to solve a subproblem.

In this interaction of child and adult there is continual shift of planning, assessments of egocentric and social ability, and communication in response to such assessments of ability. The child has a representation of the social dyad (himself and adult) including a representation of his own abilities in the sense of *individualistic ability* CAN , his abilities in the sense of *coordinative ability* CO-CAN where he utilizes the intentions of the other, and in the sense of *cooperative ability* COOP-CAN where full cooperation is required to achieve the given task (see (Werner 1990 and Werner 1994) for a semantic theory of ability including CAN , CO-CAN , and COOP-CAN).

8.2.2 *Meta-communicative interludes in children's games*

With communication the child can direct the actions of the other to set up a state such that the child can act to achieve his goals. Indeed, one often observes that a child wants help at certain points, but then wants to continue on his own. This is especially clear in cases where two or more children are playing a game. The game may require some new resource (a replacement of a ball, or finding a lost ball, water for the squirt gun, or getting some toy from the attic, etc.), but once that resource is given the adult's presence is not needed anymore and the children make a quick retreat to continue their game. Here a multi-agent interaction, the game, pauses to interact with another agent to solve a problem not solvable by the agents. When the problem is solved the original multi-agent action continues.

8.2.3 *Good-enough adult interactions*

Note, that the good-enough adult in the sense of Winnicott (Winnicott 1949) will *not* be a reactive agent that helps the child whenever there is a problem to be solved. For the adult could react and help the child at every step and even anticipate the remaining activity. That would lead to an overly dependent child unable and unwilling to solve and execute his own goals. The child abilities or action strategies are built up incrementally through action in the social physical world. An overly assisting adult is impinging his or her own needs into the process of development of the child. The other continues to be omnipotent and serve as an extension of the self rather than a distinct agent. It

mirrors the original feeding situation in the mother-child matrix.

When the adult helps the child at the child's request, the child can learn the new step or method and gradually utilize it himself by internalizing that method. Again this follows the steps of observation, building up a representation of the strategies of the other (or gaining information the other posses, e.g., the location of an object) and then internalizing those strategies by shifting the representation of the other into the boundary of the self. Pure gain in state information, such as location of an object, will increase the strategies of the agent (Werner 1991). Furthermore, the child's observations include the child's own active participation in the process. Some actions can only be accomplished with the aid of another agent. The child must learn to recognize situations that demand cooperation.

8.3 *Meta-communication and the self's dynamic structure*

It has been observed by psychoanalysts that in the interaction with the analyst the client may become more integrated or disintegrated (see (Kernberg 1967), (Kohut 1971). Indeed, Kernberg uses this as a diagnostic criterion to distinguish between borderline and schizophrenic clients. A interpretation on the structure of the self with a borderline personality will lead to a firming of the boundaries of the self. A similar structural interpretation with a schizophrenic personality will result in a regression and destabilization of the self.

How is this possible? How can the same interpretive interaction with the other lead to such different effects on the structure and function of the self?

Similarly, Kohut in a classic study of types of transference relations, observed a close dependence between the integration of the self of the client and the forms of interaction with the analyst. Thus, for example, a misunderstanding by the analyst of the emotional state of the client might lead to a shift in the organization of the self of the client. Kohut postulated that the organization of the self determines the nature of the transference (interaction type) with the other. Indeed, given such a theory the nature of the transference can then be used to determine the structure of the self of the client. More primitive transferences (mirror transference) indicated a more primitive level of self structure.

Looked at from a multi-agent perspective the communicative interaction of the self with the other can be of different types. A message (verbal or visual or tactile or olfactory) that leads to a reorganization of the self is like a meta-

interaction. The message invades areas of the self that should be private and protected in normal interactions. The pragmatic interpretation of the message or interaction sequence is no longer just at the reality level but at the self-organization level. The transformation of self that ensues can be quite confusing for both interactants.

Different kinds of meta-organizational transformations of the self can occur:

1. A message may act as a stimulus where one strategy replaces another strategy. If this is a massive switch we may speak of multiple personalities in the person.
2. A message may result in some combination of internalization and externalization. This includes 1. as a special case.

A key to understanding such interactions is that the client is in a state that includes the dyad of self and other(s). The nature of this dyad determines the force of the communicative acts (speech acts and bodily acts) in the interaction.

The typical meta-interaction as defined by the pragmatic meta-meaning of the communicative acts may be relative to the particular history and culture of the self.

8.4 *Frustrations and the structure of the self*

Also relevant here are the perceived interests of the self and the perceived nature of the power of the other to interact with the self to fulfill those self interests. For example, an immature self may perceive the interaction with the other as a nurturing dyad where the other plays the role of mother or father who is to feed or otherwise deal with the needs of the self. A failure to see the needs of the self by the actual other would lead to a state of frustration in the self and this in turn may lead to regression where the self externalizes aspects of its already sparse self leading to an even more immature organizational state of the self.

The maturity of the self will determine whether a frustrating relationship or experience results in a regression. The very same frustration may act as a stimulus to regression (a pragmatic interpretation of the event that acts on the organizational structure of the self) or it may act as a stimulus to greater autonomy (again a pragmatic interpretation of the event, but now acting to increase the competence of the self through internalization) or it may simply lead to a response by a mature self to correct the frustrating experience without

the meta-organizational pragmatic meaning of externalization (with regression) or internalization (with gain of competence). In the last case, the competence of the self and its organization remains the same. The external frustration is dealt with using strategies that are part of the repertoire of the self. The locus of these strategies, whether within the self or external to the self, is not changed.

8.5 *The false self*

In Winnicott's theory the child of an impinging mother develops a *false self*. The good enough mother reacts to child's needs whereas the impinging mother seeks out her own needs in the child, for example, going to the child when she needs emotional support. How is this expressed in terms of intentional states and internalization of interactions? The false self is a self in the sense of being a reactive self that reacts with the other in a social dialogue. However, it does not express the real interests and values of the child. Anger, for example, at a frustration may not be expressible in the interaction style of a false self. The real self is unconscious and according to Winnicott remains archaic, primitive. It is thus a simpler, more infantile intentional state. The internalization process went askew through the impinging influence of the parents. How? What does impingement mean? The metaphor of a pushing on a soft round egg wall versus a pushing out by the child is good. But how does it translate to intentions?

There must be a conflict in the child to mold it in a path against its interests. A conflict of love versus expression of anger, for example. If the child expresses anger the parent withdraws. Or if the child expresses love the parent expresses anger or withdrawal. There is then a protocol of interaction established where the child is forced to take on a role it does not really want to take on. The child may later internalize the prohibitions and reactions of the complement role.

It appears that the false self is a complex intentional state that allows the child to survive in the context of an impinging parent, but this false self does not reflect the real interests of the child. Thus there is a continual conflict between the evaluation state and the demands of the false self.

9. Self-Interest and the Boundaries of Self and Other

We have seen that internalization and externalization as well as communicative interactions dynamically form the boundary of self and other. However, power and interest also play a role here. Let us first consider how boundaries of self and other are formed from the perspective of external interactions between agents. Then we will see how such external interactions are related to the establishment of intra-psychic boundaries and the internalization process.

9.1 *Cooperation algorithms and the self*

Some time ago I investigated algorithms that generate a whole spectrum of behaviors in an agent ranging from total cooperation, to complete self-interest, to total antagonism, and even, to total self-destructivity (Werner & Reinefeld 1991). From those studies it becomes clearer that the social self can be viewed from the perspective of how the self interacts with the other. More precisely, a totally cooperative agent with no self-interest appears to have no self. Indeed, the agent functions like an extension of the other, like an arm or leg.

A system that can assume various cooperation styles will have an entire spectrum of choices of social responses ranging from total cooperation to complete antagonism. A cooperation style results from the ranking and evaluation of goals. The cooperation styles and their relationships are schematically illustrated in the picture (Fig. 1).

In the figure the quadrants are formed by two main axes, one for the self and one for the other. The attitude on any axis ranges from positive to negative (self-interested or selfdestructive, cooperative or antagonistic). For example, a totally cooperative agent with secondary self-interest will always try to achieve the goals of the other agent first before trying to achieve his own goals. The center of the circle represents complete indifference. An agent at the center has no motivation since no goal has any value at all. The algorithms make use of the agent's situation in space-time in the physical world as well as the spacetime location in the social world of the agent's own and others intentions. For a more detailed description of the various cooperation styles and their associated algorithm see (Werner 1990, Werner & Reinefeld 1990).

9.1.1 *A Totally Cooperative Robot*

For example, a totally servile robot may be experienced within the intra-

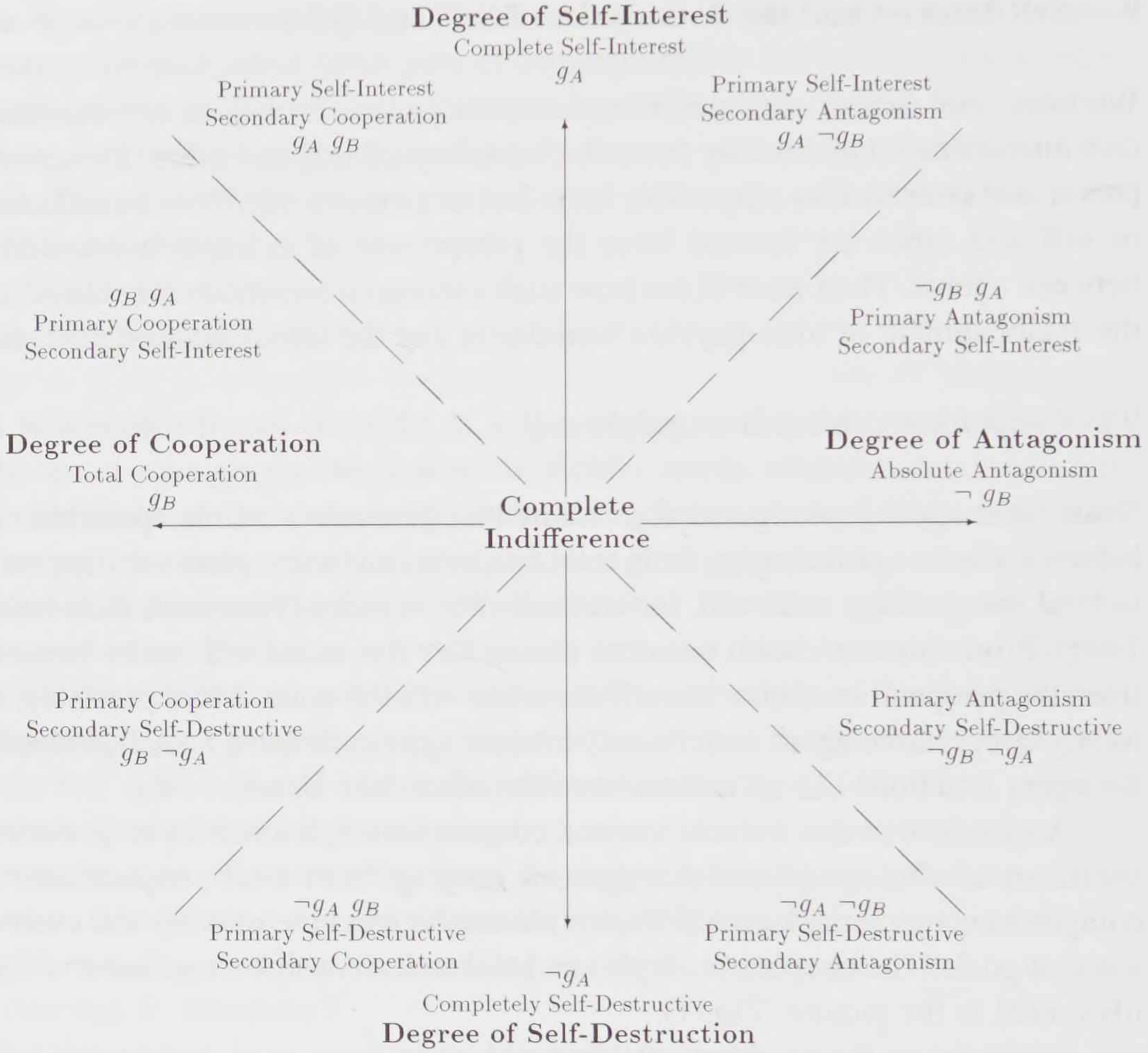


Figure 1. The Spectrum of Cooperation Styles
(© Werner 1992. All rights reserved. Reprinted with permission of author)

psychic social space as a part of an extended self of the controlling agent. For those who have seen the film Terminator II there is a scene where the heroine says that the robot in the film is like a perfect friend that is always there when one needs him. At a deeper level the robot in the film fulfills at least three unconscious wishes: First, the robot represents the perfect narcissistic object serving as an extension of the self of the actors (the mother and her son) and, by identification, the audience. Second, the robot satisfies grandiose fantasies of omnipotence by a simultaneous identification of the audience with the robot

as all powerful object. Third, the robot represents an ideal object that can do no wrong. This feeds the needs of the audience for an ideal parent. Notice, these three fantasies represent three stages of intrapsychic development of the child (Mahler 1967; 1989).

9.1.2 *Cooperation algorithms and intentions*

The point is that the cooperation algorithms act on the intentional state of the given agent taking into account the intentions of the other. Cooperation algorithms can be viewed as abstract specifications of strategies and intentions in terms of their interactions with the goals and intentions of other agents. The nature of the algorithmically specified strategy, for example, being totally cooperative with secondary self interest, defines the self in terms of its interaction with the other.

In fact, it can be a bit more complicated: Cooperation algorithms specify the response of an agent given the goals and intentions of the agent himself and not just those of other agents. Such interaction algorithms thus reflect the attitude toward self and other with respect to interests, goals and intentions. Therefore, the algorithmic specification defines the self in terms of its interaction with itself and with the other. This is not a paradoxical self-reference since it is defined dynamically, changing in time given the present intentional state of self and other. Furthermore, the more general notion of interest is involved since values or utilities for self and other are considered in addition to goals and intentions.

9.2 *Self-interest and boundaries of the self*

The boundaries of the self arise precisely with the "No!". The refusal to participate in the facilitation of the interests of the other, sets a boundary of the self when viewed as an external interaction. The greater the degree of self-interest, up to a certain point, the greater the notion of self. This is conditioned by the social needs of agents. Many goals involve others and a process of give and take over a long period of time is involved. Thus, a rationally self-interested agent will cooperate when interacting over an extended periods of time. Commitments must be met to be able to engage the other again in a cooperative interaction. Reputation becomes important.

9.3 *Evolution of the self*

In the movement of the boundaries of the self from a total encompassing grandiose self that does not differentiate between self and other to a self that recognizes its limitations, the movement to the idealized other, and then to a realistic assessment of self and other as having both common and competing interests, can be formally interpreted as a shift in the strategic boundaries of self and other with a concomitant shift in the state information and in the *interests* the agent has with respect to the world. Thus, for example, in the psychic development of the child there is an increase in the capacity of the child to assess what it can do and what it cannot do. There is an increase in abilities. And, new interests form attendant with these abilities. The child starts from a position where its interests are limited to feeding, touch, empathic interaction, and comfort. With an increase in abilities and interests the child has period of omnipotence (self locomotion, the grandiose self (Kohut 1971)), followed by a period of realization of its limitations and dependence on the other (the depressive position Mahler 1967; 1989).

9.4 *Power and the self*

The power of the self is in part defined by the cooperation of others. The psychic life of the child can be seen from a new perspective, namely, how its relations of power and interests change in the process of physical and social-psychic growth. What is grandiose may depend on what the relations of power are.

One might argue that some cultures such as in Japan do not have a "No!". Rather, in a context where each agent watches out for the interests of the other, no "No!" is needed. How does the self arise in such a social world? Some argue that such agents have no self. Rather they are more like units in a whole.

9.5 *Power and the self in anorexia nervosa*

One might view anorexia nervosa as a multi-agent interaction where there is an imbalance of power. The situation of eating reinforces power relation that is implicit in the dependency relation of receiving food. The powerless agent uses the situation of eating as a way of protesting the more general power relationship in the family. It is an act of refusal to participate in the depen-

dency relationship. In this sense it is also a striving for autonomy. Evidence for this view is that this phenomenon occurs very often in the teenage years where the self is again asserting is boundaries and its autonomy. The fact that the anorexic person must use this avenue may indicate that other avenues of self assertion where blocked off. This view concurs with the psychoanalytic work on teenagers.

The point is that a psychological phenomenon is the result of a dynamic social situation. The individuals experience that world in relationship to the power they consciously and unconsciously view themselves as possessing. Also involved may be individual goals from various levels of maturation. Thus the infantile needs of dependency may fluctuate and exacerbate more adult wishes of autonomy and assertion of self.

10. The Concept of Self

10.1 *Self as process*

The self is not a static identification like a name or a meaning; it is a way of interacting. For this reason it is more appropriate to specify the self by its way of interacting rather than by some static attributes. The self belongs more to the ontological category of process than to being. Since the self has a strategic component it does not reside in a linear four dimensional space-time. Rather it is in a multi-dimensional space of many possible universes where the self itself is constitutive of the path the actual universe will take. The self is no passive observer, but a agent that changes the universe by interacting with it either by observation (quantum mechanics) or by action (compare (Everett 1957)).

10.2 *The social self*

Since the self is inherently social, its very nature being determined by its interaction with the other, the identity of the self can be diffuse. It need not have determined boundaries like a macro (nonquantum mechanical) object. The diffusion of self in the service of social endeavors is what makes a society possible. The collective interaction of a group of agents has some similarities with the superposition of states of quantum mechanical wave-particles. The

identity of the object in quantum mechanics appears strange because it has a diffusible identity much like the human self.

10.3 *Phylogeny of the self*

The self organizational capacities of the self. One model of the self is that it is much like a plant-give it water, light and earth in the right times and amounts and it will develop on its own. The plant structures itself without assistance. The gardener has no influence of the nature or type of structure that develops. Parents with several children often take this attitude when they observe the inexplicable diversity in the personalities of their offsprings.

The fact that a coherent person or self emerges out the chaotic history of the developing self's interaction with the other, indicates, if not implies, that there exists prior structural, organizational information that guides the self's ontogeny. For the interactions with the other are circumstantial and incoherent. Ultimately, this structural information may reside in the genetic system. However, this still leaves open the nature of the prior information constituting and constraining the organization of the self, as well as leaving unanswered the questions of where, how, when, under what conditions and to what extent this information takes effect.

10.4 *Agent archetypes*

It may be that there are basic types of agents already given at birth. These may be potential or actual types that are used to interpret the social world of agents. Much like linguistic universals there may be universal agent forms that guide social experience and provide a social capacity to the agent. These universal social forms may be selected in the early and later experiences of the individual. They may provide the basic combinatoric higher order social space out of which the actual social space is selected. Certainly, there is a capacity to model social relations and interactions with others. These strategic models of interaction presuppose some representation of the agent type.

10.4.1 *A priori agent types or internalization?*

If we accept Mead, then the interaction alone determines and defines the agent type, including the subject's self. If we accept Jung, then there are basic forms for self and others within which the self and others are found. The other

alternative, is formation of self by internalization. However, internalization still presupposes a model of the other. How is the model of the other constructed?

10.4.2 *Interactionism*

One way the model of the other can come about is through an interaction with the other that leads to pragmatic operators acting on the representations of the subject to form the other by building up a strategic model (as well as an informational and evaluative model). But, this process presupposes pragmatic operators that are mature enough to interpret the interaction and to construct a model of the other. Thus the agent types are, in this case, implicitly in the pragmatic operators.

So the question whether external social interactions or prior agent archetypes enable the formation of a social self, amounts to the question of how much of the agent forming capacity is in the pragmatic operators and how much is already present in the strategic representation the agent has of self and others.

A *minimalist position* with respect to representation at birth, would claim that even the pragmatic operators that do the interpretation are learned in the interactions of child and others. The pragmatic operator itself is incrementally constructed. Still there must be a primitive meta-operator that can interpret the social world enough to build up the pragmatic operator.

10.4.3 *A library of universal agents*

The other extreme, would have an agent born with a fully formed capacity to interpret social events, i.e., a mature pragmatic operator/interpreter, and to build up agents based on a given catalogue of universal agent types or templates. For example, when the mother smiles at her infant and the infant smiles in reaction to her, does the child have a representation of a benevolent other at birth? Some, like Butterworth, argue for a significant information processing capacity (Butterworth 1998, 1996). It is possible that this capacity extends to the social processing as well.

10.4.4 *An inner social architecture*

Note, too, social organization and structure are learned, but the capacity to construct representations of social structure must also somehow be present. The *social architecture* may be only partially known by any agent in that

architecture. The question arises, is there any objective social architecture that defines the organizational structure of a group of agents? If agents have only local control knowledge and organizations form out of these local interactions, how is global organization possible? Some rough answers: Laws partially constrain. Contracts constrain. Predictions for the success of local actions in a global context can be made on the basis of local partial knowledge of global constraints.

10.4.5 *Archetypes and social knowledge*

But the point with regard to archetypes and social knowledge is that social knowledge is not just of agents and the strategies, but also of dyads and other social relations, including groups and social organizations. There may be fundamental innate social structures or forms, e.g., master-slave, friends, male-female, male-male, female-female. Or it may be that these emerge out of the basic agent types and the result of discovery and learning by the subject of the results of those typical interactions. What is known is that we have such social knowledge. The problem is how does it arise? What is given at birth? Clearly, in the animal world the mother offspring relationship is represented by both parties. There is the phenomenon of imprinting which suggests a fundamental social form or schema that is fixed or made more specific in the interaction with the actual agents in the actual context.

11. Self and the Social World

11.1 *Roles and norms*

This process of internalization can account for the gradual learning of social roles and norms. An agent acting in the context of a game, e.g., baseball, will know the roles of the other players, both on his team and the other team. The agent may utilize internalization temporarily to take on the role of another agent. But he cannot take on all roles simultaneously. He plays one of the parts. Yet, he needs to represent the other parts to do his part well in any given context. He needs to know the first baseman's job is to get the other player out before he gets to first base. His action is situated in a context of the actual state and of social roles.

11.2 *The mutual construction of the social world*

An agent can construct his world in a way consistent with his former world. He can also revise his world by contradicting it. In the case of an agent's interaction with his world, there is always a contradiction of some part of what was there before that is not there now because it has been replaced by the new result of the action of the agent.

In the case of two agents, what one agent does, affects the actions of the other agent by way of affecting the representational state of the agent. Two agents define their relation dynamically and interdependently. This can be seen clearly in the example of a man and woman who just get to know each other. Each specifies and changes their willingness to engage in various activities, accept or reject proposals, accept or reject evaluations. How far they go, depends on the willingness of both actors at each stage of their interactions. The very quality of the relationship and its organization is determined through their interaction protocols and their willingness based on their evaluation of the other in relation to their own actions. The willingness to cooperate based on evaluations and rational decision making goes beyond a mere rigid protocol of strategic response. Granted, the overall effect of the evaluations is still a protocol of interaction, but it contains variables that depend of more factors than just the messages exchanged. A woman may be uninterested in a man no matter how persuasively he is able to interact; she may not like his nose, for instance. Thus, the process of seduction depends not just on interaction protocols with rules of the type:

If A does α then respond with β

The actual resulting relationship, a social organizational dyad, cannot be determined in advance on the basis of the interaction (cooperation) styles alone. For, the partners pick their interaction style based on evaluations that go beyond the interaction style used by the other agent.

The relationship that is dynamically formed between two agents can fluctuate and change. The very interaction style used by the agents is a part of the social organization dyad. The intended effect of a social interaction style may be understood, yet rejected. The rejection immediately leads to a reassessment in both parties of the relationship. In the course of their interaction a man and woman may decide to break off their relationship. They may decide to change the status of the relationship from being lovers to being friends.

Each relationship involves rights, privileges, responsibilities and norms. Each such type of relationship involves interaction styles and forms of cooperation that are context dependent. Thus, good friends may be more cooperative for important issues, and not so cooperative for less important issues. New friends may be just the opposite. In the case of strangers in the context of the country (versus city), their may be cooperation on noncostly issues.

In the case of new acquaintances the partners discover and create their possible relationships through their interactions. Each partner discovers the relationship side of the other and each creates his own side of the relationship. Since, there is interaction, these discoveries and creations are in flux and in a process of mutual self-fulfilling and reflexive reaction.

In the case of the mother-child relationship we see a gradual change from total dependency of the child with total cooperation by the mother to more autonomy in the child and correspondingly less helpfulness by the mother. We saw how internalization furthers this process. The child also internalizes the social interaction styles of significant others. The internalization process leads to a build up of cognitive and social capacities. This in turn leads to new possible interactions between the child and his social world. And that leads to new possible social relationships.

11.3 *Social dyads*

A relationship between two people is a kind of basic organization, a social dyad.

Many patterns of interaction are fitted to one another. As one of the partners switches styles of interaction the other partner must adapt and choose a new style of interaction that fits to the change in the other. In the case of lovers the change of style may be playful and earnest at the same time. Each is trying to impress and assess the other with various strategies designed to impress and to assess.

Consider a man A and a woman W who meet each other for the first time. Let us say they are attracted to one another. This attraction may be of various forms. Neither has any information about the other except what they see. Their knowledge about the other R_A^W and R_W^A is very limited. Their relational information about one another is total uncertainty except as constrained by their own other relations and commitments. Thus, let us say the woman is married and the man is not married. The man does not know this and so

relationships such as marriage are possible. The woman, who we shall assume has a happy marriage, knows that many relationships are not possible. Thus, the two see each other in two totally different social perspectives. It will only be through their interactions that a stable mutual view of their relationship, may form.

Relational information is increased in the interaction between the parties. There may also be shifts and replacements, and not just increases.

11.4 *The fine structure of social life*

In the process of interaction with others we build up models of various types of agents and their activities in various contexts. There is a gradual learning of social situations. These involve a group of actors and a typical space of social possibilities. There is also at least one outside social context of one or more other groups. There may also be groups that are involved in the social situation. It is much like a game with players. It is also like a play in a theater with players, roles and audience. But, it goes beyond the theater, because the choices at any point are not limited to one play.

It should be noted that some might argue the theater allows just as much freedom as a game, because there is the freedom to interpret the play in many different ways. I would agree that many possible interpretations are possible for any given play, however there are even more choices in a game and not just interpretations of prescribed choices as in the case of the theater.

This debate is similar to the debate on whether agents, in general, are controlled by plans or if they interpret plans to suite their interests and goals. In the case of plans, the interpretation is somewhat analogous to the interpretation that actors and their director can give a play. Plans have the additional component that they need only roughly specify the behavior of the agent and the agent is free to make choices within that framework given the interpretation. It is like a play that does not specify all that is to be spoken and the actors are free to make up their own dialogue as they go along, in so far as it fits with the constraints the partial play gives.)

So, given a situation of conflict, for example, where an enemy is attacking, the person being attacked can model that as a social dyad with various options.

1. The agent can run away and a chase situation of hide and seek may occur. So the agent in evaluating this option would have to evaluate the attacker's seeking ability and his escaping and hiding ability.

2. The agent may stay and fight. He may then seek a weapon or tool to increase his fighting ability.
3. If language is available, the agent may attempt negotiation through communication.
4. Within a larger social context, the agent may seek help from a third party, who forms a coalition with the agent against the attacker.
5. Also in a larger social context that contains agents with more power than either of the two agents in the social situation, the agent may seek help or arbitration from a third party who protects him from the enemy. In cases of arbitration, this option may result in the arbitrator deciding against the agent and joining the attacker. Then the agent has a coalition of two agents against him.

As the agent gains more abilities and social knowledge, this primitive social schema for attack may become a quite sophisticated social space for dealing with social conflict. It reminds one a bit of Schank and Abelson's scripts (Schank & Abelson 1977). But there is a difference here that the agents have choices, so it is more like a game. It also more than a game because social roles, social power, social organizations are all involved in the social space.

12. A Formal Theory of the Social World

12.1 *Social space time*

The social world is as real as the physical world. Indeed for the agent it is just as important. What an agent can do, his very abilities depend on the social space that surrounds the agent and on the agent's knowledge of that social space. Actually, if we look at the logic of ability, it is a modal operator that depends on the agent's information about the *physical space-time* world and it depends on the agent's knowledge of his *social space-time world*. For the semantics of "*A* can α" where α denotes some goal or event in the future, contains the agent's information state $I_A(H^t)$, the information the agent has in his physical world H at time t. It also contains the agent's own intentions $S_A(I(H^t)$ in his world H at time t given his state information I. Finally, the semantics also includes the agent's knowledge of his social world SW, namely S_A^{SW}.

The social world *SW* consists of the intentional world of the society that surrounds the agent. It has a multi-dimensional structure that includes time and possible spacetime world developments. This social space-time world is real in the sense that the agent can have an incorrect representation of that world. Certainly, the agent has only a partial representation of the social world. Thus, the agent's representation of the world in social space time S_A^{SW} allows many different possible intentional structures. If the agent's representation of the social world is correct, then the *actual social world* (in intentional space time) *ASW* will be one of the possible intentional social worlds allowed by S_A^{SW}. Thus, in the possibility formalism, $ASW \in S_A^{SW}$.

Note, that a social world has a history consisting of social states much like a physical possible world. We can thus distinguish a social history from the state of the social world at a given time. In fact, a *social history SH* is a sequence of social world states.

12.2 *Social space and social worlds*

Social states are highly modal. By that I mean that they are highly intentional and contain a tremendous amount of uncertainty due to the freedom of choice which we assume the individual agents possess. The term social world suggests a physical state or linear history. So, when I want to emphasize the potentiality, uncertainty, possibility and intentionality of social worlds, I will use the term social space. For space is really a representation of possibilities. A 3D *physical space* represents all possible movements in that space. A *social space* represents all possible developments (choices of agents) as constrained by the intentions of those agents. A given *social history* just describes one possible path a social world.

One might think, since the social world ultimately reduces to choices which are actions and which reduce to events in physical space-time, that the social world is only a mental fiction that disappears as soon as the choices are made. All that is left after the choices of all the agents is a physical world state. That attitude has motivated and misguided formalisms that treat actions as operators on physical world states. And, it has resulted in a misplaced emphasis on the physical effects of choices.

Each choice of an agent may effect the social world. A simple action a, for example, may be observed by another agent B. That perception of a is interpreted by a pragmatic operator a_α that operates on B's information state I_B

including B's information about A's information state I_B^A. This in turn may effect his own intentional state S_B, including his knowledge of A's intentions S_B^A. For example, B may wake up at night and observe an agent A trying to wedge open B's window. Meta observations (observing the observer) may result in cascades of social interactions. For example, Agent A observing agent B observing A doing α will influence A's intentions and future actions. Thus, the thief realizes he has been discovered and he may activate his fight or flight decision strategy.

A social history of social world states exhibits potentiality and intentionality at each state. In a history of physical world states each choice of an agent or the physical world as agent, results in a new physical world state. The agent operates on those states. The actual resulting state is not determined by the action of one agent, but by the interaction of the actions of all the agents (including the world agent).

In the social world, each choice may result in a new intentional structure of the multi-agent world. Indeed, this intentional structure may change just from mental choices (or computation, reasoning, preferences, new wants, new goals, new intentions). To the extent an action (physical, message, gesture) is observed it changes the information state of the observing agents. And, this in turn may change the intentions of the observing agent. The outside physical world state need not change as the social world changes. This is why in the traditional view of action as being an operator on physical states fails to capture the intentional, pragmatic meaning of action in the context of a social world. For example, at a party, a man may make a fool of himself, the skillful socialite will not acknowledge this, but her intentional state and that of her guests has changed. The entire real social world for the drunkard has changed even though the physical world gives no evidence of this and even though the drunkard's representation of the social world has not changed.

12.3 *Misrepresentation of the social world*

An agent can misrepresent the social world. Usually, we do misrepresent some aspects of the social world. The process of adjustment of a misrepresented social space to a space more in line with social reality can now be described in more detail. Here, I am using represented social space to mean a typical representation of the social world that indicates the possibilities, options and abilities available to the agents in that social space.

The agent carries with him his social space. The man who has been in prison for 10 years will, at least at first, carry with him his social space as it was in the prison and project it into his interactions in the world. *Socialization* is the adjustment of an agent with his given social space to a new social space. More formally, it is a gradual transformation of the prison social space to the actual social space.

Social spaces are not seen directly but are discovered and formed dynamically in the interaction of the agent with the social world. Furthermore, since the agent always carries with him a given social space, all interactions the agent is involved in are interpreted and mediated in terms of his active social space. I say *active social space*, because the agent may have several social spaces. Much like the meaning of a message depends on the representational state R of the agent, so the meaning of the social situation and the agent's reaction depends on his given, active social space. When a message α is received by an agent with information state I, the interpretation of α depends on I since the pragmatic operator $Prag(\alpha)$ acts on the given information state I. The only possible way to change the social space of an agent is if the other agents do things that are not expected, i.e., not included or not permitted in the social space of the subject. Or, the agent may do some new action not permissible (but still) within the social space and get a reaction not predicted by the social space. One of the pleasures of travel is that it can unhinge one's normal social space. Emotions appear to be relative to an active social space.

Note, in the case of infantile reactions, the subject carries with him social spaces that are activated by certain contexts. These social spaces may contain traumatic situations for the subject, e.g., hurt, shame, guilt, or fear of danger. And, once active, the subject interprets all actions of the other agent's in terms of that social space. The Oedipal triangle is just one instance of a social space. The early mother child dyad is an even more primitive social space. The early interaction with playmates of the toddler are also social spaces. The play with other significant adults, e.g., a grandfather or aunt, are also part of the social space. The relations with siblings and in relation to the significant adults are also very important parts of the social space. As the child gains more capacities these social spaces also change and grow. But they can also freeze in the case of trauma. Social spaces may be repressed and become unconscious for the subject, and at the same time be active. The subject may feel the emotions of the *ur* social space and not be conscious of that space. The space is in the background, a process running in parallel with the conscious social space.

13. Macro and Micro Social Laws: Max Weber

Marxism and sociological positivism attempted to explain social activity exclusively through external, objective laws that did not consider the subjective intentions of the agents. For Weber (Weber 1958, 1997) the meaning of action is intersubjective and not the result of the system or social whole. In Marxism change occurs through external, systemic, economic forces where the individual action is ineffectual. For Weber change results from the actions of agents. Human agents are motivated to act in specific ways. Their cognitive state is essential for determining their actions. For example, the spirit of capitalism with its origins in Calvinism involves specific values: 'of systematic self-control' and the 'systematic use of time', thrift, saving, proving their worth in this world. Calvinism demanded adherence to these maxims in everyday conduct.

At the theoretical level it is a dispute between global control through universal, objective laws and local control through invariancies across the intentional states of individual agents. Note, the *values* are maxims for intentional behavior. Behavior and actions are evaluated and not just objects. This evaluation results in their adoption by the agent who accepts those values. Globally it increases the probability of an agent acting in accordance with those values. Locally the agent's actions are constrained by those values. Locally the rationality of the agent, his information state, his goals and intentions, together with his evaluations determines his actions. Globally, the universality of a property of the intentional state (control or plan state) of the individual agents, results in the emergence of global, systemic properties. These global redundancies thus result in forms of organization or social patterns.

A key issue is how far does the local control state and its properties, if multiplied across many agents, determine global agent interactions? Note, also that the local agents need not all be in the same intentional state. Indeed, most cooperative activity, or competitive activity, involves the assumption of complementary roles and styles of interaction. Such styles of interaction can be very specific, and detailed to attain the coordination that is necessary for achieving social patterns. This sort of cooperation can occur between groups of agents as well.

Weber's definition of 'social relationships ... denote the behavior of a plurality of actors in so far as, in its meaningful content, the action of each

takes account of that of others and is oriented in these terms.' The definition can be formalized in terms of the intentional states that agents consider in making their decisions. Thus, the logic of CAN , CO-CAN and COOP-CAN is relevant here. Note, the definition is not detailed enough to characterize social relationships based on cooperation styles since the definition does not include any constitutive intentional structures or attitudes. Instead, the definition depicts individual, rational agents making decisions on the basis of the intentions of other agents. It gives no account of complex social relationships where action is constituted by complex dynamic, complementary strategic states. Such complex social actions cannot be reduced to one-step rational decision making even if it includes reflection about the other agent.

Weber's concept of *spirit* can be defined as a common set of values, including evaluations of behavior (not just individual, unit actions but a whole strategy of action), that lead to general properties of the intentional states of the agents assuming or assimilating those values. These general properties are constraints on those intentions and which result in constituting, defining and controlling agent actions. The claim by Weber is that this spirit can lead to particular forms of human or agent organized action. So, universal properties in a set of agents leads to global properties of the set of agents as a whole.

An invariance in the properties of the intentions of a set of agents is a basic kind of global property of that set of agents. Such an invariance can be used to delimit a subset of the set of all agents.

In so far as the agents themselves can recognize an invariant property of agents, they can use it to determine membership, in any given case, in that subset of agents. However, the agent's determination of membership based on properties that are intentional and that, therefore, can be feigned, may also be based on other interests that distort the official criteria for membership. Thus, for example, a member of a program committee of a conference may be incompetent to really review the incoming papers, but because of political connections that member may be tolerated. And, if, as often happens, the incompetents gain the upper hand, competent people are actually excluded from membership because they are a threat. So too, if the corrupt are in the majority or in all the right places, the noncorrupt are systematically excluded or undermined.

14. Conclusion

The theory of social psychology presents a novel paradigm for the investigation of psychodynamic phenomena. The focus of this new paradigm is not information as it is in cognitive science but intention and value. In social agent psychology information plays more of a subsidiary role. Not only can we expect to understand human social interactions in a more detailed way, but also the psychodynamic interactions of artificial social beings. Like it or not we will be faced with serious problems of interaction as artificial social agents become more commonplace.

We have tried to provide an answer as to how autonomous agents that are initially independent can become social beings. The process of socialization was explored through the psychological process of internalization. Externalization was seen to be complementary to internalization. The psychological process of internalization is made possible through the interactions an agent has with his social world. This interaction is mediated by gestures, language and the observation of, and the participation in activity. The theory of internalization and socialization presented here can apply to quite simple agents as well as more complex agents like human beings. Our theory of internalization makes role acquisition possible. This in turn makes organized activity between groups of agents possible. The continuation of a social system from generation to generation is thus explained through the continual renewal and birth of the social self.

References

Balint, M. 1968. *The Basic Fault: Therapeutic Aspects of Regression*. Travistock, London.

Butterworth, G. 1998. "What is special about pointing in babies?". *The development of sensory, motor and cognitive capacities in early infancy: From the perspective of cognition*. Simion, F., & Butterworth, G. (Editors), Psychology Press/Erlbaum Taylor & Francis, Hove, England.

Butterworth, G. 1998. "A developmental-ecological perspective on Strawson's 'the self'". *Journal of Consciousness Studies*, Vol. 5(2), 132-140, 1998.

Butterworth, G. 1996. "The origins of language and thought in early childhood". *Handbook of Human Symbolic Evolution*, Lock, A., et al., (Editor). Oxford University Press.

Everett, H. 1957. "Relative State Formulation of Quantum Mechanics", *Reviews of Modern Physics*, vol. 29, 454-462.

Freud, A. 1946. *The Ego and the Mechanisms of Defense*. International Universities Press, New York.

Grice, H.P. 1968. "Utterer's Meaning, Sentence-Meaning, and Word Meaning". *Foundations of Language*, Nymegen, Netherlands, Vol. 4, 225-242.

Grice, H.P. 1957. "Logic and Conversation". *Syntax and Semantics*, Los Angeles, CA, Vol. 3.

Habermas, J. 1981. *Theorie des Kommunikativen Handelns*. Suhrkamp, Frankfurt am Main.

Jacobson, E. 1971. *Depression*. International Universities Press, New York.

Jung, C.G. 1923. *Psychological Types*. Harcourt, Brace, New York.

Kernberg, O. 1967. "Borderline Personality Organization". *Journal of American Psychoanalytic Association*, vol. 15, pages 641-685.

Kernberg, O. 1970. "Factors in the Psychoanalytic Treatment of Narcissistic Personalities". *Journal of American Psychoanalytic Association*, vol. 18, pages 55-58.

Kernberg, O. 1995. *Borderline Conditions and Pathological Narcissism.* Jason Aronson, April 1995.

Kohut, H. 1971. *The Analysis of Self; A Systematic Approach to the Psychoanalytic Treatment of Narcissistic Personality Disorders,* International Universities Press, June 1971.

Kohut, H. 1976. *The Restoration of Self.* International Universities Press, October 1976.

Lacan, J. 1966. *Écrits*. Seuil, 1966.

Lacan, J. 1982. *Écrits*. A. Sheridan (translator), W.W. Norman and Company.

Mahler, M. 1967. "A Study of the Separation Individuation Process and its Possible Application to Borderline Phenomena in the Psychoanalytic Situation". *Psychoanalytic Study of the Child*, vol. 26, pages 403-424.

Mahler, M. 1968. *On Human Symbiosis and the Vicissitudes of Individuation in Infantile Psychosis*. International Universities Press, June 1968.

Mahler, M., Bergman, A., and Pine, F. 1989. *The Psychological Birth of the Human Infant: Symbiosis and Individuation*. Basic Books, September 1989.

Mead, G.H. 1967. *Mind, Self and Society from the Standpoint of a Social Behaviorist*. Charles. W. Morris (editor), University of Chicago Press, August 1967.

Miller, A. 1979. *Das Drama des begabten Kindes und die Suche nach dem wahren Selbst*. Suhrkamp Verlag, Frankfurt-am-Main.

Schank, R.C., & Abelson, R. 1977. *Scripts, Plans, Goals, & Understanding; An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, Inc.

Searle, H.F. 1979. *Countertransference and Related Subjects*. International Universities Press, New York.

Shannon, C.E. 1948. "The Mathematical Theory of Communication". *Bell Syst, Techn. Journ.*, vol. 27, pages 379-423; 623-656.

Spitz, R. 1966. *First Year of Life: A Psychoanalytic Study of Normal and Deviant Development of Object Relations*. International Universities Press, June 1966.

Von Neumann, J. and Morgenstern, O. 1947. *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Vygotsky, L.S. 1978. *Mind in Society*. M. Cole, et al. (Editors). Harvard Univ Press, London.

Weber, Max M. 1958. *From Max Weber: Essays in Sociology*. H. H. Gerth (editor), Oxford Univ Press, December 1958.

Weber, Max M. 1997. *Theory of Social and Economic Organization*. Talcott Parsons (Introduction), Free Press, July 1997.

Werner, E. 1988a. "Toward a Theory of Communication and Cooperation for Multiagent Planning". *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 2nd Conference*. Morgan Kaufmann Publishers, pp. 129-142.

Werner, E. 1988b. "A Formal Computational Semantics and Pragmatics of Speech Acts", COLING-88. *Proceedings of the 12th International Conference on Computational Linguistics*. Budapest, Hungary.

Werner, E. 1988c. "Social Intentions". *Panel on Interaction Among Intelligent Agents, Proceedings of the 8th European Conference on Artificial Intelligence*. Munich, Germany, pp. 719-723.

Werner, E. 1989. "Cooperating Agents: A Unified Theory of Communication and Social Structure". *Distributed Artificial Intelligence*, Vol. 2, M. Huhns & L. Gasser (eds.). Morgan Kaufmann and Pitman Publishers, London, pp. 3-36.

Werner, E. 1990. "What Can Agents Do Together? A Semantics of Cooperative Ability". *ECAI-90, Proceedings of the 9th European Conference on Artificial Intelligence*. Stockholm, Sweden, Pitman Publishers, pp. 694-701.

Werner, E. 1991a. "A Unified View of Information, Intention, and Ability". *Decentralized AI, Vol. II*, Y. Demazeau & J-P. Müller (eds.). Elsevier Science Publishers (North Holland).

Werner, E. and Reinefeld, A. 1991b. "Distributed Algorithms for Cooperating Agents". *DAI-90, Proceedings of the 10th International Workshop on Distributed Artificial Intelligence*. Bandera, Texas.

Werner, E., & Demazeau, Y. 1992. *Decentralized Artificial Intelligence III*. North- Holland, Amsterdam.

Werner, E. 1994. "Logical Foundations of Artificial Intelligence". *Foundations of Artificial Intelligence*, G. Ohare, et.al. (eds). Wiley, London.

Winnicott, D.W. 1949. "Hate in the Countertransference". *Inter. Journal Psychoanalysis*, vol. 30, pp. 69-74.

Winnicott, D.W. 1965. *Maturational Processes and the Facilitating Environment: Studies in the Theory of Emotional Development*. International Universities Press, June 1965.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford, Basil Blackwell.

Wittgenstein, L. 1958. *The Blue and Brown Books*. Blackwell, Oxford.

CHAPTER ELEVEN

Computational Embodiment

Agents as Constructed Complex Systems

Christopher Landauer and Kirstie L. Bellman

1. Introduction

We are trying to construct intelligent agents that can both operate autonomously and cooperate with each other in complex environments. The two questions considered in this Chapter are: (1) "What are the properties that makes an agent social?" (Bellman 1997b) (Dautenhahn 1998), and (2) "How can we construct flexible systems that will allow us to develop and test increasingly sophisticated concepts of intelligent agents in a social environment?" (Bellman & Landauer 1996a). We generally regard answers to the first question as requirements and answers to the second question as implementation designs that are intended to satisfy those requirements.

In our opinion, there are several properties that must occur for an agent to be properly called "social": it must observe other agents' behaviors, and make internal explanations of them. It must have a kind of awareness of itself in context, which means a viewpoint that allows it to reason about its own behavior in the context in which it operates (we do not presume that this implies any kind of "consciousness"). It must also be able to encounter and interact with other agents, cooperatively or otherwise, with some notion of a group and of goals larger than one individual (this is also part of the social context).

In order to act with any of these social properties, we believe that agents require a high degree of autonomy (e.g., they need the representations and capabilities to respond in accordance with internally-generated goals). Of course, this autonomy may at first be limited to the tactical level (how to do it), with a much more limited strategic capability (what to do). Since an artificial agent is always doing *something* for *someone*, whether that someone is a human user or another agent or computer program, all artificial agents must be social in our sense to be called agents. Otherwise, they are just computer programs with some interesting logical properties and behaviors.

A quick look at the brains of living systems will convince one of how much infrastructure is necessary for a computational system to implement any kind of behavior, even the simplest behaviors common to all biological systems. Furthermore, certain kinds of social behavior seem to be emergent, requiring no internal intelligence at all (Kennedy 1999). This possibility adds to the difficulty of characterizing the capabilities that agents exhibiting social behavior will have to have, since it implies that at least some of these social properties have to be thought of as "belonging" not so much to the individual as to the collective, and not so much to a collective *per se* as to the dynamic properties of collectives within an "ecological niche". These hard questions will require a great deal of empirical observations and engineering experimentation in building different kinds of agents. Hence, we welcome the recent proliferation of agents that we see in the field today. However, we hope that in this Chapter we can also encourage the necessary discussion for building a theoretical basis to these systems.

In this chapter, we describe an approach to constructing autonomous computational agents that is based on theoretical work on the organization of structures that underlie language and movement processes in biological systems, and on the structure of *constructed* complex systems mediated or integrated by software. We are not only interested in the usual control problems involved in developing systems that can carry out planned behaviors or respond adaptively to unanticipated events, but also in instrumentation to find out what is going on inside and outside the system, in on-line management of system resources for efficiency, in negotiation among system resources, and in system architectures that will help us design, build, use, maintain, and analyze such systems.

Parts of this chapter have appeared earlier (Bellman & Landauer 1997d) (Landauer & Bellman 1999c), some in a very limited-distribution context (Landauer & Bellman 1997c).

1.1 *Autonomy*

In our opinion, there are really only two classes of (difficult) requirements for effective autonomy: robustness and timeliness. Robustness means graceful degradation in increasingly hostile environments. Timeliness means that situations are recognized "well enough" and "soon enough", and that "good enough" actions are taken "soon enough". Both of these are forms of adaptive behavior, and neither one of these implies any *optimization* necessarily.

The basis for all computational autonomy lies in knowing what is out there (appropriate sensors and models in useful conceptual spaces), organizing how to deal with it (control sharing mechanisms), determining what to do (appropriate reasoning and reflection), and doing it soon enough (appropriate effectors and real-time processing). We have left these and most other implementation issues out of this chapter, because we are more concerned here with fundamental theoretical issues, namely, how interesting autonomous behavior is possible at all (a rock is autonomous, but not interesting). If some progress can be made on this last question, the implementations will follow. To start looking at this issue, we look at what observing living systems has to teach us.

1.2 *Infrastructure for complex systems*

Our research in complex systems has shown the importance of infrastructure (Landauer & Bellman 1996a, 1996c), that is, explicit components and activities of the system whose function is to help organize the rest (Allen & Garlan 1994), (Landauer & Bellman 1995a), (Hayes-Roth et al. 1995), (Shaw 1995) and (Shaw & Garlan 1996). No matter what kinds of computational models are used, the system will need infrastructure which supports complex interactions. In Section 3, we briefly describe our "Wrapping" approach to intelligent integration infrastructure for constructed complex systems, which provides a natural means for incorporating adaptation and other processes as computational resources, using the same adaptation and flexibility mechanisms for both the agent's behavior within an environment and for the construction of the agents, other processes and tools, and the operating environment itself.

2. Taking a Hint from Biology

It is instructive to compare computational systems with biological systems, since those are the only ones we know that have interesting complexities of behavior (Bellman & Walter 1984). From the very smallest and simplest organisms, biological systems have adaptive processes that provide appropriate actions with sufficient flexibility to adjust to changes in their environment. In this section, we describe some of the qualities of biological systems that permit such robust behavior and which we would like to incorporate into agent research. Our purpose is to improve the capabilities of computer-based agents, so that they can work with humans to perform increasingly difficult cognitive tasks, such as finding relevant information, filtering and refining the presentation of information for our purposes, and organizing and executing the sequence of command steps for such machines as factory product lines or autonomous land vehicles. These last tasks of dealing with the planning, command, control, and execution of behaviors blur the already thin line between "robots" and "agents". We believe that this thin line should disappear completely, so that the growing field of agents research can make more use of the hard-won lessons from biological research and robotics about what it means to carry out intelligent functioning within ANY world (be it abstract or physical). Four lessons that we have drawn from these fields are that "intelligent agency" requires:

1. developing "ecological niches" or portable contexts for computational agents;
2. creating artificial "embodiment" for abstract agents;
3. creating entities with "social" behaviors; and
4. developing capabilities for growth and adaptivity of behaviors.

Before we expand on each of these points below, we want to emphasize here that each of these directions is being considered in current agent research (Maes 1990b), but often only implicitly. If there is one thing that Artificial Intelligence has taught all computer scientists, it is the power of making hidden assumptions or viewpoints explicit and hence processable. Assumptions about contexts, about interactions with other agents, and about how and why the agents' behaviors should be done a certain way are all being built into the agents. Instead, we are suggesting that by placing the agents within an explicit context (with its own processes for being examined and modified), by giving agents a definite "shape" and body within this context, and by giving

agents a means of socially responding to other agents in explicit and observable ways, we can strengthen the agents' performances by creating new and stronger information with which to evaluate our ideas of intelligent functionality by agents. We believe that ecological niches, embodiment, and social intelligence are all good starts to building agents with a much richer repertoire of mechanisms for adaptation and growth.

Especially the grounding of agent behaviors within an observable context, gives us the opportunity to build active, generative processes so that an agent can monitor its own behavior, change it, catalog it, and learn it. But in order to capitalize on the creation of such processes, the agent needs (1) architectures that allow it to pull in new types of processing resources, and (2) self-reflective capabilities. Later on, we briefly describe our Wrapping approach to dynamic infrastructure, which is one way of providing the flexibility and infrastructure that supports these two goals.

2.1 *Ecological niches for agents*

In nature, the capabilities and characteristics of any animal are strongly related to the environment within which it lives. The context within which the animal must perform includes the physical world, the other species that exist within that area, the history of its own species within that environment, and its immediate social interactions with others of its own kind. Biologists continue to struggle with understanding the kinds of powerful principles that lead to the extraordinary amount of specialization and variety that has resulted (Braitenberg 1977).

Creating abstract information handlers has led to a pervasive myth in the Information Sciences about "general purpose intelligence" and "general purpose processors". One of the gaps that occurs between theoretical computer science and applications is here: in applications, one must have "agent" programs that have access to specific knowledge, that do rather precise things with that knowledge, and that present it in very specific ways. These specializations are often built into these agent programs. We want to draw them out, examine them, and create a new way of organizing them for explicit processing.

We need to take a lesson from biology here: there are no such things as "general purpose fish". They are all recognizably fish, and hence there are many general aspects that we can use in creating architectures and standard processes for "fishness". However, fish are specialized by living in saltwater

or fresh; living near the top sunlit waters or being bottom dwellers; by surviving near fierce predators or by having few competitors; by living in schools or developing other protective mechanisms; and so forth. Eventually, we want to enumerate the corresponding "ecologically" important distinctions for "agents", e.g., do they live on the Internet or in a digital library; do they have to fight for our attention or do they have dedicated lines or other access to us; do they work as part of a temporary alliance of known (or harder, unknown) resources, or are they a simple logic or filter component in a larger process; and so forth.

The advantages of building explicit contexts for our agents are clear from a research point of view. Instead of trying to align our "agents" and their behaviors to a hidden or implicit context, we can start to represent and manipulate "niches" explicitly, study the mechanisms by which our agents relate to that kind of context, study the behaviors of our agents within the explicit contexts, and begin to build principles of how we distribute information and processing capabilities between the agent and its operating environment.

The first issue to consider in the interaction between a system and its environment is that there are certain constraints over which the system has no control at all. Some behaviors "in the world" affect an agent regardless of its intentions (e.g., time, gravity). By the "physics of the world", we mean the permanent constraints on interactions between the agent and its environment, both static and dynamic. In so far as we get these constraints wrong in the design of our systems, the systems will have trouble accounting for their effects in their own planning.

Multi-User Domains (MUDs) are an interesting new kind of groupware that incorporates people into the program. MUDs have become enormously popular as games and as educational support tools over the last few years, because they get the human interactions right in some fundamental sense, and because they engage our sense of "place". MUD clients and servers are easy to obtain and run (most servers and clients are free), but they usually only provide text worlds; there is little interaction with existing tools that are outside the MUD; though some have construction languages that allow complex programming, it is the usual kind of programming; and it is not very easy to access large volumes of information. We have been exploring the use of MUDs as an ecological niche for agents (Landauer & Bellman 1996b).

2.2 *Embodiment*

The next hard-won lesson from biology and robotics is that intelligent functionality needs a body, e.g., it needs a shape that defines and limits its capabilities and scope. In animals, bodies reflect both the short-term and long-term history of a species' adaptation to its ecological niche. In animals, bodies reflect the physical constraints of the environments, the historical and immediate interactions, choices, and methods of dealing with the environment and others. The embodiment is the grounding of the agent and its means of relating to its environment.

Two other critical issues emerge from considering bodies. The first is that an embodied agent reminds us that an agent always exists within its world, and that it is always doing something within that world. Even "idling" is a behavior, the result of which can be quite unpleasant for prey in the real world, or simply annoying to a user waiting for performance. In one sense, researchers have already given their agents "bodies" in the sense that they have given the cluster of computer programs that comprises their agent a set of interfaces (to a real or virtual environment, to instrumentation, sensors, and effectors, to other agents, to the developer and to the users). As in the preceding section, we believe that these interfaces should be made more explicit, observable, and hence analyzable, by elaborating and integrating such interfaces into a "body". The embodiment is an explicit representation of the "Integration Concept" for the "whole" animal. This integration concept provides a unity to the whole system. It is what turns a bag of parts into a coherently operating individual entity. A body has many parts, but works as one entity, integrated across disparate resources (unless it is sick or dysfunctional).

With this "body", we must account for the causal connection to and from the environment, that is, the effects of events in the environment on the agent, and the effects of agent activities on external objects. Hence, the agent can learn of these effects or cause others only through the specific interface between the agent and the environment. Although, having the appropriate sensors to collect information from the environment is often emphasized in agent work, it is just as critical for agents to have the right effectors. The interface must have the right effectors to make actions occur: when the agents act, there must be some sort of impact on the environment (even if that is only changes of the agent's state within that environment). In many models of social interactions, there are missing steps that explain exactly how the inputs

and decisions of the agents are mapped into actions with consequences or results in the environment. The idea of embodiment requires one to define these steps in sufficient depth. This embodiment must also include the means by which the agent will be able to monitor when they have performed some action and its consequences; this feedback is critical in the use of tools by the agent and in its ability to monitor other activities for their the proper effect. One big difference and advantage of artificial systems over biological ones is that in artificial environments we can also instrument the environments in new ways that will allow us to monitor and analyze behaviors and interactions previously not possible (Bellman 1999a).

The second is that the body is a barrier; we cannot reach into the embodied system and turn switches when we want something to happen (Bellman 1996a). Commands are only suggestions, since we cannot force any particular kind of behavior on the system. We can provide compelling reasons, and try to arrange that we have programmed the system to respond appropriately, but we can not enforce it. This is one of the main differences we see between computer programs that we want to call agents and ordinary computer programs.

In our opinion, some of the most important aspects of autonomous systems revolve around this notion of "embodiment", that is, the connection of the computation to an entity in the physical world. Embodiment in this sense has several essential properties:

- a sense of environment: the environment has properties and contains tools, objects, and other agents, and makes certain actions available. The environment starts where the "body" ends.
- a sense of presence: there is a causal connection to and from the environment, so that the use of tools and other activities have a direct impact on it. This sense is an awareness of ones "body" within a given environment.
- a sense of time: it is important for the agent to remember the history of what has gone before, to recognize certain kinds of event patterns, and to make predictions. The sense of time is deeply dependent upon the rate at which one's body moves, e.g., something is fast when we can not move as fast and slow when we can move a lot faster (Uexkull 1934).
- a sense of place: what happens and what exists take place in some locality, which has a notion of geography, connectivity or connection, distance, and space. This is one's label or name for, or just recognition of, the environment where one's body is.

- a sense of will: autonomous means self-governing, so an autonomous system must generate its own goals, and exhibit a sense of purpose and intention. A body is the extent of what the agent directly controls with its will.
- a sense of self: an autonomous system needs to have a notion of its own resources, abilities, and internal state. The body is one key part of defining what self and non-self is. There are many provocative stories of the effects of violations of the boundaries of one's body (Sacks 1985).
- a sense of perspective: in order to make effective choices and recognize opportunities and threats, an autonomous system needs a viewpoint of itself in its environment. The body's placement within an environment, the extent of its sensors and effectors, and its history are key determinants of "perspective" (Bellman 1999a).

2.3 *Social intelligence*

Part of the context for any animal is the presence of other individuals, both competing species and members of its own species. If we imagine the jungle of the Internet, we can well envision the situation in which one's own agents will have to deal not only with well-behaved conspecific agents, but all sorts of tools with different degrees of intelligence and behavior (Dautenhahn 1999). They may even have to deal with malicious software or rogue agents.

Furthermore, even if we have agents that deal only with carefully selected others, we need to develop much better ideas of what it means to communicate and behave cooperatively among agents. Many social scientists, Cyberneticians (such as Gordon Pask), and now a few computer scientists (Bond & Gasser 1988), (Hewitt & Inman 1991) have been pointing out for some time that cooperative behaviors — shared goals, work, understandings, communications — do not occur inside a participant alone. In addition to the interpretations of the acts or symbols within each participant alone, there is also an act of negotiation and agreement among participants as to the meaning of acts (including speech acts). These mutually-defined meanings are necessary in order for the group to interact even on the simplest levels, e.g., to determine what constitutes the desired results of even individually determined actions, or what constitutes that an action has been "done" in order for another agent to proceed. Clearly these "meanings" can be built in (and in fact usually are in most agent interactions, in the form of stopping rules or predefined precondi-

tions for further activity), but eventually we want to have agents that are "smart" enough to act more autonomously and adaptively.

A key aspect of being able to negotiate among agents turns out to be the ability to share a common context with other agents, to observe the other behaviors within that context, and to construct explanations of those behaviors, so that they may be integrated by an agent into a local view of the other agents (Dautenhahn 1999). The notion of niche allows us to create places where the behaviors of an agent are visible and interpretable both by itself and by others. In an immediate practical sense, this increases the "bandwidth" of information that others can process about the agent; instead of just relying on messages and self-report, other agents can observe and infer usable information from the agent's behavior. Eventually, we want to use such behavioristic information to form the basis for a sharable semantics — a real semantics, grounded in shared experience within this niche — for human-agent communication (Bellman 1997b).

2.4 *Growth, adaptation, and development*

One of the main motivations for developing intelligent agents is to build systems that have more flexible and interesting ways of changing their behaviors under the right circumstances. The behavior of animals is so sophisticated in this regards that for many centuries most scientists and philosophers believed that there could never be an adequate explanation of any human or animal behavior. In rebellion to the dogma of this early vitalism, this generation has erred often in the opposite direction, applying overly simplified mechanistic explanations to behaviors. Even the simplest of animals (single-celled animals for example) show formidable adaptive capabilities beyond our current best machine processes (Bellman & Goldberg 1984).

That is, to be a living creature in a real environment on this planet already requires some sophisticated adaptive capabilities. Although we do not have time to discuss adaptation in biological systems in depth, we would like to bring into this discussion two key theoretical concepts, developed in several early papers (Bellman & Goldberg 1984), (Bellman & Walter 1984) that are highly relevant to agent architectures and infrastructure: (1) the difference between generative processes and knowledge-intensive processes, and (2) the second author's concept of "variation spaces" from (Bellman 1979).

First, we start with observation of *generative processes* in animal sys-

tems. By "generative processes", we mean two related things: processes that create the "information" used by the animal in its processing, and also the processes that allow an animal to produce behavior in a variable way. For example, sensors (both of the external environment or one's movement within that environment, such as proprioceptive ones) provide streams of data that, when combined with such cognitive capabilities as memory, recognition, and reasoning, result in information that can then be utilized by an animal to plan actions, evaluate them, and learn from them. For some time now ethologists, biologists and psychologists have been approaching the very difficult issues of how such streams of data are multiply represented (from different sensory and cognitive processes), how different sensory and cognitive processes are integrated with each other and with ongoing movement and activities, and how such "information" is generated at all or utilized by an animal system in an ongoing and graceful fashion.

In addition, there are a number of mechanisms in animal systems that allow them to gracefully generate a large number of behaviors that are, at once, both typical and specialized to the needs of the moment. In both language and movement, one sees behavior that seems to obey highly consistent rules, but at the same time, permits the generation of a very large set of instances specialized to the momentary needs of the animal. For example, the way that a lizard holds its limb reflects at the same time several behaviors (e.g., starting to lean forwards towards some food and yet crouching, slightly afraid, in front of the observing human), the slope of the ground on which it is standing, its history of injuries and health, and so forth (Bellman 1979).

Such ability to generate such large families of related, but specialized, configurations is partly the result of dozens of mechanisms underlying the nervous system and the body, such as having nerves and muscles that permit graded movement, visual and proprioceptive feedback, the physical linkages among muscles, joints, and limb bones and the cognitive processing that permits the planning and coordination of several behavioral goals. One key part of this is that in animal systems there are such mechanisms at many different levels of the system, from cellular through behavioral. Also, in both meanings of generative processes noted above (as the creation of information and as the creation of variable behavior), one is, in fact, drawing from a large set of underlying mechanisms that permit rule-governed and patterned behavior, the tailoring and specialization of patterns to immediate constraints, and the creation of information that appears to be encoded for later use and

modifiecation. The difference between the higher cognitive behaviors and the sensori-motor behaviors, the location and use of the encoded information will of course be the matter of research for many years to come.

In contrast to this, information systems are much simpler. Information systems do not yet have the interesting diversity of such generative mechanisms or the constraints inherent in an ecological niche and the physics of a real environment. Instead, they are knowledge-intensive, that is, they must depend for the most part on pre-existing stores of both information and rules in order to create responses to new input conditions. Artificial intelligence has made some real gains in creating better reasoning processes and more clever search strategies in finding and evaluating applicable rules (Luger 1994). However, this only allows the enumeration of logical combinations (such as one sees in a decision table), rather than the adaptive generation of animal behavior, that changes in a continuous fashion along multiple dimensions, and responds at the same time to multiple types of constraints. Providing a niche, a social setting, and a body as noted above are a few ways in which we believe we can move towards the invention of more interesting and generative processing.

Bellman's concept of "variation spaces" is a way to characterize the solution space within which a system can adapt (Bellman 1979). That is, in order for any system to adapt, it first depends upon a set of underlying mechanisms and structures that allows it to vary in meaningful ways. Meaningful here refers to what is important to that animal population within that habitat and to the individual animal within its lifetime and history of experiences. Animal systems change through a number of different means over several distinct time scales. An individual animal is the member of a species whose adaptations are measured over thousands of generations and whose results are encoded in its genome and in the accretion of the artifacts from thousands of lives within a habitat (resulting in such diverse ecological niches as coral reefs and human cities). The changes in species appear as changes in the look (color, size, material), the sensory systems of, the behavior of, and the habitats of individual animals.

An individual animal also grows and develops over its lifetime. Some adaptive mechanisms come into play only at certain times within this life. The changes of an animal over its lifetime can be as dramatic as the changes appearing in species (some juvenile forms live in entirely different habitats and have different limbs and sensory organs), or less apparent, being seen as

size, coloration, and behaviors. Lastly, an individual animal responds with changes in its behavior, both immediately and over longer periods of time through learning. Hence a human responds both to immediate requirements (a traffic signal, a telephone call, the response to a question in a conversation) and learns new behaviors over longer periods (to walk and talk, to read and write, to play tennis and so forth). These changes in humans and animals are often seen as size of musculature, speed of response, "knowing" the right way home or through a maze, and so forth. The point here is that none of these changes would be possible unless there first existed a set of mechanisms that allowed the animal to vary in any of these dimensions. Generative processes are one important source of such "controlled sources of variation". There are many others, such as the musculature that allows our limbs to move in many directions, eyeballs that allow us to look in different directions, a cortex that allows us to correlate diverse sensory impressions, or to associate different ideas, our metabolism that permits weight gain and loss etc. All of the different mechanisms that permit variations in the animal system can then become potentially the means by which an animal system can adapt to some new circumstances.

Taken together, one can think of these mechanisms that underlie controlled sources of variation in a system as setting up the boundaries of a multi-dimensional variation space. That variation space becomes the solution space within which one can generate or describe any behavior by that organism or system. This basic set of mechanisms needs to be complemented by a second set of adaptive mechanisms that act upon this variation space to create the appropriate behavior for a given situation. For example, this second set of mechanisms might represent overlays of constraints upon the variation space, pruning the available solutions, or they might determine a usual ordered set of configurations ("cutting a usual path" through that space), or they might help create "attractors" within this space that help the system remain within a limited set of solutions despite small perturbations.

We have been applying the notion of variation spaces to agent architectures by explicitly analyzing where we want the flexibilities in agent behaviors (and their testbeds) and the means by which we can create the variation (by either explicitly storing different cases, as in a rulebase, or through the development of computational methods that take us a step closer to more generative processes). To summarize, there must be processes that construct variation spaces within which adaptation can take place, and decision pro-

cesses that can make adaptive choices within those spaces. Furthermore, variation spaces can change over time. We want some changes to be directed by the history of the agent in its environment (this is not just what it has "learned", but rather the results of the information we keep on the results of agents in given environments used to refine the agent architectures), some by the dynamics of the environmental niche (changes in the environment can "reshape" the solution space), and some by the "intention" of the agent (its state, capabilities, knowledge, priorities and goals also "reshape" the variation space).

Furthermore, since the *purpose* of adaptation in systems is to allow them to be effective in uncertain dynamic environments, this argues for complex heterogeneous systems, and that regardless of the individual algorithms used for adaptation, the rest of the software framework must have certain properties for adaptation to work. Adaptation requires flexibility; every flexibility must have a corresponding coordination mechanism, and every such pairing can use different methods for adaptation. In fact, every such pairing *must* use different methods for adaptation, at least superficially, since their context and scope are different. This argument leads directly to the requirement for multiplicity in *ALL* aspects of the system, including the basic adaptation algorithms. Many kinds of computing systems have been called adaptive, but in most of them, the processes that control the adaptation are built-in, separately from the processes subject to adaptation, so that they are not themselves available for flexible reuse and modification during software development.

The absence of generative processes means that the system developers must build explicitly all of the "sources of controlled variation", the behaviors, and parts of the system, and hence severely limits our ability to construct large systems with many types of adaptive and coordinative behavior. One of our philosophical principles for system architecture is the analogue of the "controlled sources of variation" mentioned above: every flexibility we want in a constructed complex system must have a corresponding coordinative mechanism that manages the flexibility. We have made some progress in defining essential features of complex computing systems (Landauer & Bellman 1996a, 1996b, 1997d) (Bellman 1997b).

Adaptation of external interaction behavior (input collection, motion, and communication), using internal representations of that behavior and its effects, is important for real-world embedding, because it lets the system adjust the dynamic range of its sensors to the current environmental characteristics.

Adaptation of internal representations (notations, interpreters, and other processes) is the same process, applied to the internal processing in the system. Finally, adaptation of the agent to its environment involves a number of special processes, such as exploration, discovery, learning, prediction, and planning, which we discuss briefly below. An earlier version of this section appeared in (Landauer & Bellman 1997c, 1999c).

3. Reflective Infrastructure for Agents

In this section, we summarize our Wrapping approach to integration infrastructure in constructed complex systems, our use of what we have called *knowledge-based polymorphism* to construct computationally reflective systems. It is only a short overview of the Wrapping approach; many more details, including the basic algorithms used in our Problem Managers, can be found elsewhere (Landauer & Bellman 1996a, 1996b, 1997f) (and references therein).

Our Wrapping approach to constructing heterogeneous software and hardware environments is based on two key complementary parts: (1) explicit, machine-processable descriptions of all software, hardware, and other computational resources in a system, and (2) active integration processes that select, adapt, and combine these resources for particular problems.

In the Wrapping approach, we not only emphasize meta-knowledge about the uses of computational resources, and brokering and mediation of all component interactions (all critical concepts, as seen increasingly in other approaches), but we also regard as equally important the special resources for organizing and processing this information in a flexible and evolvable fashion.

The Wrapping approach, because it wraps all of its resources, even the active integration processes (Landauer & Bellman 1993a, 1996a), results in systems that are Computationally Reflective (Smith 1984, 1986) (Maes & Nardi 1988) (Kiczales et al. 1991). That is, a system organized in this way has a machine-processable model of itself; the Wrapping resources and their interactions allow, in essence, a simulation of the entire system to be contained within the system. This allows sophisticated instrumentation and adaptive processing. It is this ability of the system to analyze and modify its own behavior that provides the power and flexibility of resource use. These ideas have proven to be useful, even when implemented and applied in informal and

ad hoc ways (Miller & Quilici 1992) (Bellman & Reinhardt 1993).

The wrapping theory has four fundamental properties that we regard as essential (Landauer & Bellman 1996a, 1997a):

1. EVERY part of a system architecture is a *resource* that provides an *information service*, including programs, data, user interfaces, architecture and interconnection models, and everything else.
2. EVERY activity in a system is *problem study*, (i.e., all activities *apply* a resource to a *posed problem*), including user interactions, information requests and announcements within the system, service or processing requests, and all other processing behavior. We therefore specifically separate the problem to be studied from the resources that might study it.
3. *Wrapping Knowledge Bases (WKBs)* contain *wrappings*, which are explicit machine-processable descriptions of all of the resources and how they can be applied to problems to support what we have called the *Intelligent User Support (IUS)* functions (Bellman 1991b):
   - *Selection* (which resources to apply to a problem),
   - *Assembly* (how to let them work together),
   - *Integration* (when and why they should work together),
   - *Adaptation* (how to adjust them to work on the problem), and
   - *Explanation* (why certain resources were or will be used).

   Wrappings contain much more than "how" to use a resource. They also include both qualitative and quantitative information to help decide "when" it is appropriate to use it, "why" you might want to, and "whether" it can be used in this current problem and context.
4. *Problem Managers (PMs)*, including the *Study Managers (SMs)* and the *Coordination Manager (CM)*, are the active integration processes: algorithms that use the wrapping descriptions to collect and select resources to apply to problems. They use implicit invocation, both context and problem dependent, to choose and organize resources. The PMs are also resources, and they are also wrapped.

The Wrappings information and processes form expert interfaces to all of the different ways to use resources in a heterogeneous system that are known to the system (Landauer 1990a).

The most important conceptual simplifications that the Wrapping approach brings to integration are the uniformities of the first two features: the uniformity of treating everything in the system as resources, and the uniformity of treating everything that happens in the system as a problem study. The

most important algorithmic simplification is the reflection provided by treating the PMs as resources themselves: we explicitly make the entire system reflective by considering these programs that process the Wrappings to be resources also, and wrapping them, so that all of our integration support processes apply to themselves, too. It is this ability of the system to analyze its own behavior that provides some of the power and flexibility of resource use, and that we believe is essential for effective autonomy in computing systems (Landauer & Bellman 1999c).

4. Prospects for Autonomy: Summary and Conclusions

We believe that autonomous systems must be complex systems, with a great repertoire of possible behaviors and processes for selecting them, fallback choices to account for incorrect situation estimation, and quick partial solutions to reduce decision time. All activity is situated, strongly dependent on context, and there need to be different decision processes in different situations. These conditions require integration of many different kinds of models of external and internal environment, architecture, behavior, and the system itself, in their respective contexts. The models are both predictive and empirical ones, and they require continual validation processes. This flexibility requires an architecture in which many parts of the system are infrastructure, organizing other parts of the system to identify and address problems, and monitoring their behavior (Hayes-Roth et al. 1995).

We believe that with proper attention to niches, embodiment, social behaviors, and architectures supporting adaptive and reflective behaviors, we can build agents that are more useful partners in an increasingly complex information environment. In summary, we consider adaptation to have two somewhat different aspects: one is the construction and selection of "variation spaces" within which the adaptations may occur, and the other is the selection of actions within those spaces. Biological systems seem to be able to generate both of these kinds of processes as needed, at least to a limited extent. In order that the system can respond to a wide dynamic range of possible environmental conditions, a very broad range of potential behaviors must be available to the system.

In other papers, we show how the necessary flexibility and coordination can be implemented, using our Wrapping approach to integration (Landauer &

Bellman 1997f), and how they allow us to think of a common Interaction Space (Landauer & Bellman 1996b), in which humans and computer programs can interact in more interesting ways (Bellman & Landauer 1996a). Since Wrappings allow many special case processes to be combined with general case methods, and Multi-User Virtual Environments (MUVEs) allow multiple programs to be connected together with multiple humans as users, all in the same environment, we have taken the Wrappings architecture and applied it to the MUVE servers to define these Interaction Spaces.

Unlike a formal mathematical space or even the usual homogeneous simulation system, part of the strength of a MUVE is its ability to become the common meeting ground for a variety of different types of symbol systems and processing capabilities. These different symbol systems and processes occur in a variety of forms within the MUVE. Information and processing capabilities can be "packaged" as "agents", who often interact with human users in natural language, and can freely move and act in the same way as a human user within the MUVE; as an "object" that is manipulated by human and agent users within the MUVE; or as part of the "setting", e.g., the description, capabilities, and "physics" of one of the many "places" within a MUVE.

The packaging of some process to become an object, an agent, or part of a setting in a MUVE hides, as any good interface does, many details about how that process works. The MUVE gives the appearance of a uniform world of objects, actors, and places, all acting within the common "physics" of a setting, and seen and heard in the same way. This is a reasonably successful and good strategy. However, if one looks one level deeper, this common meeting ground is also the theoretical meeting ground for how one will communicate among different types of formal systems. Here then is a new hopeful tack for a hard traditional problem of integrating different formal systems.

We are applying Wrappings to agent architectures by building "agent carriers" for programs (arbitrary computer processes can be embodied as characters within the virtual worlds of the MUVEs); they carry programs into MUVEs as spacecraft carry payloads into space. We are currently studying the information structures required of such programs, from control modularity (separation of external interfaces from command and response streams) to software disintegration (identifying and separating computational resources from a large system), semi-automatic construction and automatic elaboration

of the Wrappings, and other software engineering issues (Landauer & Bellman 1996a, 1996c).

With these MUVEs or interaction spaces, we hope to study new ways of integrating the activities and capabilities of both human and artificial agents, with the capabilities of the virtual habitats and objects within it. The computer-mediated environments permit all the tool-to-tool, human- using-tool, and human-to-human interactions to be recorded in a potentially machine-processable form. We also will be able to experiment with not only different protocols, but with the role of "common experiences" within the same worlds on the coordination of behavior and communications.

To construct the Wrappings for interaction spaces, we need to define the information services provided by the various programs, and the kinds of information and protocols used. This information is generally available for interesting programs, though it is not usually in an explicit machine-processable form. We believe that these new MUVE servers will be excellent testbeds in which protocols, man-machine interfaces, coordination among multiple users and resources, and the processing of contextual information may be studied in a more explicit and enriched fashion, and because of these properties, that they can become an important testbed for studying increasingly sophisticated program behavior.

Intelligence is much more mundane than programs playing hard symbolic games or proving theorems. It is also much more remarkable. The ability for an organism to exist in the world, to move around, interact with its environment, and make experience-dependent decisions, is one of the deepest mysteries of life. These things that most living things can do (at least for a while) are much more interesting and difficult than the things only humans can do, and the things most humans can do "without thinking" are at least as interesting and difficult as the things a few humans can only do after much study or training. We have just begun to approach the challenges of understanding how humans can cooperatively use their intelligence. But an even greater one lies before us: what can humans complemented by intelligent agents do? If the variety already appearing in Virtual Worlds built up by hundreds of humans over only a few years, and inhabited by primitive robots is any indicator, the answers will surprise even the most visionary among us.

References

Allen, Robert & David Garlan. 1994. Formal Connectors. CMU-CS-94–115.

Bellman, K. 1979. *The Conflict Behavior of the Lizard, Sceloporus Occidentalis, and Its Implication for the Organization of Motor Behavior*. Ph.D. Dissertation, UCSD.

Bellman, Kirstie L. 1993. "An Approach to Integrating and Creating Flexible Software Environments Supporting the Design of Complex Systems", pp. 1101–1105 in Proceedings of WSC'91: The 1991 Winter Simulation Conference, 8–11 December 1991, Phoenix, Arizona (1991); revised version in Kirstie L. Bellman, Christopher Landauer, "Flexible Software Environments Supporting the Design of Complex Systems", Proceedings of the Artificial Intelligence in Logistics Meeting, 8–10 March 1993, Williamsburg, Va., American Defense Preparedness Association.

Bellman, Kirstie L. 1996. "When Intelligence is in Control", pp. 10–12 in Intelligent Systems: A Semiotic Perspective, Proceedings of the 1996 International Multidisciplinary Conference, Volume I: Theoretical Semiotics, Workshop on Biologically Inspired Complex Systems, 20–23 October 1996, Gaithersburg, Maryland.

Bellman, Kirstie L. 1997. "Sharing Work, Experience, Interpretation, and maybe even Meanings Between Natural and Artificial Agents" (invited paper), pp. 4127–4132 (Vol. 5) in Proceedings of SMC'97: the 1997 IEEE International Conference on Systems, Man, and Cybernetics, 12–15 October 1997, Orlando, Florida.

Bellman, Kirstie L. 1999. "Towards a Place Theory with Virtual Worlds", pp. 17–21 in Proceedings of VWsim'99: Virtual Worlds and Simulation Conference, SCS Western Multi-Conference, 18–20 January 1999, San Francisco, SCS.

Bellman, Kirstie L. & Lou Goldberg. 1984. "Common Origin of Linguistic and Movement Abilities", American Journal of Physiology, Volume 246, pp. R915–R921.

Bellman, Kirstie L. & Christopher Landauer. 1997. "A Note on Improving the Capabilities of Software Agents", pp. 512– 513 in Proceedings AA'97: The First International Conference on Autonomous Agents, 5–8 February 1997, Marina Del Rey.

Bellman, Kirstie L. & Christopher Landauer. 1997. "Computational Embodiment: Biological Considerations", pp. 422– 427 in A. M. Meystel (ed.), Proceedings of ISAS'97: The 1997 International Conference on Intelligent Systems and Semiotics: A Learning Perspective, 22–25 September 1997, NIST, Gaithersburg, Maryland.

Bellman, Kirstie L. & Captain Al Reinhardt. 1993. USAF, "Debris Analysis Workstation: A Modelling Environment for Studies on Space Debris", Proceedings of the First European Conference on Space Debris, 5–7 April 1993, Darmstadt, Germany.

Bellman, Kirstie L. & Donald O. Walter. 1984. "Biological Processing", American Journal of Physiology, Volume 246, pp. R860–R867.

Bond, Alan H. & Les Gasser (eds.). 1988. Readings in Distributed Artificial Intelligence, Morgan Kaufmann.

Braitenberg, Valentino . 1977. On the Texture of Brains, Springer-Verlag.

Dautenhahn, Kerstin. 1998. "The art of designing socially intelligent agents: science, fiction, and the human in the loop", Applied Artificial Intelligence, Volume 1, No. 7, pp. 573-617.

Dautenhahn, Kerstin. 1999. "Socially Situated Life-Like Agents", pp. 191–196 Proceedings VWsim'99: The 1999 Virtual Worlds and Simulation Conference, part of WMC'99: The 1999 SCS Western Multi-Conference, 18–20 January, San Francisco 1999), pp. 191-196.

Davis, Paul K. & Richard Hillestad (eds.). 1993. Proceedings of DARPA Variable-Resolution Modeling Conference, 5–6 May 1992, Herndon, Virginia, Conference Proceedings CF-103–DARPA, RAND Corp.

Hayes-Roth, Barbara, Karl Pfleger, Philippe Lalanda, Philippe Morignot & Marka Balabanovic. 1995. "A Domain-Specific Software Architecture for Adaptive Intelligent Systems", IEEE Transactions on Software Engineering, Vol. SE-21, No. 4, pp. 288–301.

Hewitt, Carl & Jeff Inman. 1991. "DAI Betwixt and Between: From "Intelligent Agents" to Open Systems Science", IEEE Transactions on Systems, Man, and Cybernetics, Volume SMC-21, No. 6, pp. 1409–1419.

Kennedy, Jim. 1999. "Artificial Sociocognition: Cultures of Belief in Populations of EleMentals", pp. 50–55 Proceedings VWsim'99: The 1999 Virtual Worlds and Simulation Conference, part of WMC'99: The 1999 SCS Western Multi-Conference, 18–20 January, San Francisco.

Kiczales, Gregor, Jim des Rivifieres & Daniel G. Bobrow. 1991. The Art of the Meta-Object Protocol, MIT Press.

Landauer, Christopher. 1990. "Wrapping Mathematical Tools", pp. 261–266 in Proceedings of the 1990 SCS Eastern MultiConference, 23–26 April 1990, Nashville, Tennessee, Simulation Series, Volume 22, No. 3, SCS (1990); also pp. 415–419 in Proceedings of Interface'90: The 22nd Symposium on the Interface (between Computer Science and Statistics), 17–19 May 1990, East Lansing, Michigan.

Landauer, Christopher & Kirstie L. Bellman. 1993. "The Role of Self-Referential Logics in a Software Architecture Using Wrappings", Proceedings of ISS'93: the 3rd Irvine Software Symposium, 30 April 1993, U. C. Irvine, California.

Landauer, Christopher & Kirstie L. Bellman. 1995. "The Organization and Active Processing of Meta-Knowledge for Large-Scale Dynamic Integration", pp. 149–160 in Proceedings 10th IEEE International Symposium on Intelligent Control, Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems, 27–30 August 1995, Monterey.

Landauer, Christopher & Kirstie L. Bellman. 1996a. "Knowledge-Based Integration Infrastructure for Complex Systems", International Journal of Intelligent Control and Systems, Volume 1, No. 1, pp. 133–153.

Landauer, Christopher & Kirstie L. Bellman. 1996b. "Integration Systems and Interaction Spaces", pp. 161–178 in Proceedings of the First International Workshop on Frontiers of Combining Systems, 26–29 March 1995, Munich.

Landauer, Christopher & Kirstie L. Bellman. 1996c. "Constructed Complex Systems: Issues, Architectures and Wrappings", pp. 233–238 in Proceedings EMCSR 96: Thirteenth European Meeting on Cybernetics and Systems Research, Symposium on Complex Systems Analysis and Design, 9–12 April 1996, Vienna.

Landauer, Christopher & Kirstie L. Bellman. 1997a. "Model-Based Simulation Design with Wrappings", pp. 169–174 in Proceedings of OOS'97: Object Oriented Simulation Conference, Part of WMC'97: The 1997 SCS Western Multi-Conference, 12–15 January, Phoenix, SCS International.

Landauer, Christopher & Kirstie L. Bellman. 1997b. "Computational Embodiment: Constructing Autonomous Software Systems", pp. 42–54 in Judith A. Lombardi (ed.), Continuing the Conversation: Dialogues in Cybernetics, Volume I, Proceedings of the

1997 ASC Conference, American Society for Cybernetics, 8–12 March 1997, U. Illinois (1997); poster summary in (Bellman Landauer 1997a); revised version in (Landauer Bellman 1999).

Landauer, Christopher & Kirstie L. Bellman. 1997c. "Computational Embodiment: Software Architectures", pp. 205–210 in A. M. Meystel (ed.), Proceedings of ISAS'97: The 1997 International Conference on Intelligent Systems and Semiotics: A Learning Perspective, 22–25 September 1997, NIST, Gaithersburg, Maryland.

Landauer, Christopher & Kirstie L. Bellman. 1998. "Wrappings for Software Development", pp. 420–429 in 31st Hawaii Conference on System Sciences, Volume III: Emerging Technologies, 6–9 January 1998, Kona, Hawaii.

Landauer, Christopher & Kirstie L. Bellman.1999. "Computational Embodiment: Constructing Autonomous Software Systems", pp. 131–168 Cybernetics and Systems Journal, Volume 30, No. 2.

Luger, George F. with Peder Johnson, Carl Stern, Jean E. Newman, and Ronald Yeo.1994. Cognitive Science: The Science of Intelligent Systems, Academic Press.

Maes, Pattie (ed.). 1993. Special Issues of Robotics and Autonomous Systems, Volume 6, Nos. 1 and 2 (June 1990); reprinted as Pattie Maes (ed.), Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back, MIT / Elsevier.

Maes, Pattie & D. Nardi (eds.).1988. Meta-Level Architectures and Reflection, Proceedings of the Workshop on Meta-Level Architectures and Reflection, 27–30 October 1986, Alghero, Italy, North- Holland.

Miller, Lawrence H. & Alex Quilici. 1992. "A Knowledge- Based Approach to Encouraging Reuse of Simulation and Modeling Programs", in Proceedings of SEKE'92: The Fourth International Conference on Software Engineering and Knowledge Engineering, IEEE Press.

Sacks, Oliver.1985. The Man Who Mistook His Wife for a Hat, Summit Books.

Meystel, Alex.1995. Semiotic Modeling and Situation Analysis: An Introduction, AdRem, Inc.

Shaw, Mary.1995. "Patterns for Software Architectures" Chapter 24, pp. 453–462 in James O. Coplien, Douglas C. Schmidt (eds.), "Pattern Languages of Program Design", Proceedings PLoP'94: First Annual Conference on Pattern Languages of Programs, August 1994, Monticello, Illinois.

Shaw, Mary & David Garlan. 1996. Software Architecture: Perspectives on an Emerging Discipline, Prentice-Hall.

Smith, Brian Cantwell.1984. "Reflection and Semantics in LISP", Proceedings of PoPL'84: The Eleventh Annual ACM Symposium on Principles of Programming Languages, 15–18 January 1984, Salt Lake City.

Smith, Brian Cantwell.1986."Varieties of Self-Reference", in Joseph Y. Halpern (ed.), Proceedings of TARK'86: Theoretical Aspects of Reasoning about Knowledge, 19–22 March 1986, Monterey, California.

Uexkull, Jakob von.1957. "A Stroll Through the World of Animals and Men", (1934); reprinted in Instinctive Behavior: the Development of a Modern Concept, Translated by Claire Schiller, International Universities Press.

CHAPTER TWELVE

Are We Having Fun Yet? Using Social Agents in Social Domains

Leonard N. Foner

1. Introduction

It is sometimes desirable to create a social agent — one that people can interact with as if it was, at least in part, another person — because it makes a particular application easier to program, easier to use, or more entertaining to use. But when exactly might one want to do this, how would one go about it, how does human cognition either help or hinder this, and how can one evaluate the results? This paper attempts to answer these questions.

We employ a standard methodology of *qualitative analysis* (Maxwell 1996) (Miles and Huberman 1994) (Wolcott 1990) here, in which we examine a single case, in depth, as a model for one possible application domain and implementation strategy. Such an examination can suggest general principles which can be validated by construction and evaluation of additional test cases.

The overall conclusions of this study are:

- The proper domain is critical.
- The tendency of people to anthropomorphize can be used to advantage.
- Simple implementations are often sufficient.
- User expectations can make or break a social agent.

Or, to put this even more succinctly: If you can choose an appropriate domain, a social agent is often the most natural, easy-to-program, and easy-to-interact choice for how to structure the computation.

This chapter is organized as follows. First, we describe the concept of a

Turing test. Next, we describe ELIZA, which motivated much further exploration of Turing tests, and PARRY, which was itself motivated by previous research in Turing tests.

We then spend the bulk of the paper in an examination of the domain of MUDs, and the general capabilities of Julia, the agent under consideration. As part of this discussion, we investigate the simplicity of her implementation. We then look at some sociology, and evaluate how those who interact with her feel about the interaction. Finally, we use these lessons to draw some conclusions.

2. The Turing Test

Much of the discussion in this paper concerns the *Turing test,* a famous (and famously-misunderstood) element in computer lore. What exactly *is* the Turing test? And, for that matter, who was Turing?

Alan Turing (1912–1954), a British citizen, was and is one of the most famous computer scientists ever. He originated the *Turing machine*, had a critical role in Allied cryptanalysis during World War II, and also invented the *Turing test.*

The *Turing machine,* first described in 1936, is his description of the mathematics behind a typical stored-program computer, and is the singular model of how conventional digital computers work. Numerous mathematical proofs of the properties of what can, and cannot, be computed via a Turing machine similarly demonstrate what can, and cannot, be computed by any stored-program digital computer.

Our purposes here, however, concern the *Turing test.* As used in almost all discussions of natural-language processing by computer scientists, it is taken to mean *a test to determine whether a program can imitate a human being in typewritten conversation.* The procedure to be employed consists of a human in one room, a set of judges in another, and a computer somewhere, all connected via some sort of text-only telecommunications system. The judges, who do not know which communications link is connected to the human and which to the computer, ask both a series of questions, and attempt to decide from their respective answers which is the human and which is the computer. Presumably, either the human, the computer, or both, attempt to deceive the judges. If the judges cannot tell which party is which, the computer is said to have *passed a Turing test.*

This was not Turing's original description of his famous test (Turing 1950); instead, he proposed *telling male from female* — in other words, the two parties were one male and one female, rather than one human and one computer.[1] However, for our purposes here, we shall take the better-known, human/computer version of the Turing test as our benchmark.

3. Some Important Early Examples

When one talks to any computer scientist about natural language processing, or the Turing Test, ELIZA is often the example that leaps to mind.

3.1 *ELIZA*

ELIZA (Weizenbaum 1966, 1983) was originally written by Joseph Weizenbaum in 1966, while at the MIT AI Lab, as an experiment in natural-language processing. It simulates a Rogerian psychotherapy session — the user types in a single sentence, and the program responds with another sentence. The program did simple pronoun substitution (turning "I" into "you" and so forth), and often quoted part of the user's reply back in the response. For example, a typical interchange might look like this, with a strict alternation between the user and the program:

User: I'm angry.
ELIZA: Are you angry often?
User: Yes.
ELIZA: I see... Well, why do you say this is so?
User: Maybe it's because of my mother.
ELIZA: Are you sure that is the real reason?
User: Why do you ask?
ELIZA: Does it bother you that I ask?
User: I guess not.
ELIZA: Maybe your mother has something to do with this.

Since Rogerian therapy is generally completely undirected, and since the program also tended to answer questions with questions, the aimless nature of the resulting conversation was unusually well-tolerated by users. Indeed, the rapidity with which users imbued human qualities to ELIZA — often going out of their way to keep their responses within the envelope to which ELIZA would respond sensibly so as not to shatter the illusion — greatly disturbed

Weizenbaum, who commented, "What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people." (Weizenbaum 1976).

Yet the program's attraction persists. It was translated into BASIC and published in an early hobbyist-computer magazine in 1977 (North and Shrager 1977).[2] A version ran on the MIT AI Lab PDP-20 in the mid-1980s as the "operator sysjob," and unsuspecting users who try to send messages to the mainframe's operator — there was no actual human operator in attendance- would instead have their statements answered by ELIZA, with often amusing results. (All such interactions were logged, and reading the resulting transcripts was great, if cruel, fun — an exchange such as "Can you please mount a tape for me?" answered by, "Is it because of your mother that you want me to mount a tape for you?" could only lead to disaster.)

It was later translated into GNU Emacs Lisp and became a standard part of the distribution of GNU Emacs around 1990 (type M-x doctor to interact with it). It has been translated into innumerable other languages, such as Perl. Clearly, ELIZA has captured the imagination of many in the field.

Mauldin (Mauldin 1994) has commented that ELIZA makes use of a number of simple tricks to achieve vaguely convincing performance:

- ELIZA answers most utterances with a question. This tends to elicit further conversation, since most users like to talk about themselves, and the program appears to be willing to listen.
- ELIZA often incorporates substrings of the user's responses in its replies (with suitable pronoun substitution), which makes it appears that the program is able to follow the conversation.
- Its use of Rogerian psychotherapy means that it never really says anything declaratively, which means it will never contradict itself later.

3.2 *PARRY*

An important later example of this sort of program is Colby's PARRY (Colby 1971), which, unlike ELIZA, had explicitly psychological goals — it was written to decide whether psychotherapy could be said to understand paranoid thinking. If a judging panel could not successfully differentiate PARRY's responses from those of a true clinically paranoid patient, then it could be argued that we *did* understand paranoia. In other words, if we can model the *external appearance* of paranoia, perhaps our model of how a paranoid patient thinks is itself accurate.

This is the same sort of reasoning used in the commonly-accepted view of the Turing test — if a computer cannot be distinguished from a human, then we must have captured computationally something about what it means to think like a human. While such an assertion is open to endless debate (and, indeed, there is a tremendous literature arguing the point), the programs that have been produced in an attempt to answer the question have been quite interesting. ELIZA, PARRY, and Julia (see below) are some of the results.

PARRY demonstrated a surprisingly accurate grasp of human paranoid behavior. Testing revealed that trained psychotherapists could not differentiate PARRY's responses in a transcript from those of paranoid patients.[3]

PARRY uses ELIZA's bag of tricks, and then some (Mauldin 1994):

- It can admit that it doesn't know the answer to a question.
- It can change the level of the conversation, e.g., "Why do you ask?"
- It keeps returning to its "pet topic" — small stories about the Mafia-which it tries to tell in a fixed order, much like many paranoid patients.
- It occasionally introduces a new topic, much as a real person would.

4. An Introduction to MUDs and Julia

We now turn our attention to the star of this paper, *Julia*.

Julia (Mauldin 1994) is a MUD robot, originally developed to try to pass Turing tests. A MUD is a text-only, multiperson, virtual reality (Bruckman 1993) (Curtis 1992). MUDs started out as games, but they have not remained so. For example, MUDs have been used as the basis for a computer-supported collaborative work *[CSCW]* environment for technical professionals in user-support and system administration tasks (Evard 1993).

Before we begin, a note about pronouns. Many of the examples of interaction with Julia are from the perspective of the author of this paper; in such cases, it is often more readable and more compact to use a first-person pronoun, whether singular or plural, in place of "the author." In addition, since Julia adamantly insists that she is female, we refer to the program here as "she."

We start off this description of Julia by situating her in her environment, since the domain of Julia's operation is a critical feature. Julia functions by interacting with other users of the MUD, so we show transcripts of many interactions with her as examples. [An even larger set of interactions with Julia is presented in (Foner 1997).] We will investigate details of her imple-

mentation later, but it should be noted that a simple description of the implementation, as in (Mauldin 1994), gives very little "feel" for what it is like to interact with her outside of the strictures of a formal Turing test.

4.1 *Introduction*

Julia connects as any human player on the MUD would, via a telnet connection — she does *not* run on the MUD server itself. A player would gain access to her numerous talents in exactly the same way that a player would talk to any other player — by either speaking or gesturing in the same room, or by paging from somewhere else:

> >page julia
> You sent your summons to Julia.
> Julia pages from Stevi's Kitchen: 'I'm on my way to you, Lenny.'
> Julia is briefly visible through the mist.
> Julia says, 'I was called here by Lenny.'
> You say, 'julia?'
> Julia nods to Lenny.
> […]
> Julia whispers, 'Excuse me, Xerxes is paging me from Connie's place.'
> Julia walks south to the airship landing field.
> Julia has left.

This is a continuing thread in our discussion of Julia: players interact with her *as if she was another player*. MUDs blur the distinction between players and 'bots (MUD slang for a software robot). This is a great strength of both MUDs and Julia, and the deliberate blurriness of the boundaries both enhances Julia's effectiveness and makes her operation possible.

One of Julia's most useful talents is to keep maps of the interconnections between rooms. The way to get Julia to help you navigate is to ask her a question, as usual.

> You say, 'julia, how do I get to electronics den?'
> Julia says, 'From here, go south, 7, airlock, 9, up, down'
> Minotaur says, 'Julia, how do you get to the gold sector?'
> Julia says, 'From here, go highport, gold, library'

Since Julia keeps track of various statistics about rooms (rooms being very important objects in any MUD), she can answer certain questions that might be important to a mudder, including how many rooms or players there are, which rooms are "noisiest" due to conversation, where one can go from any

room, and will even make suggestions:

> TheHighMage says, 'Julia, I'm bored. Where should I go?'
> Julia says, 'You should go see gilded door, TheHighMage.'
> Julia says, 'From here, go present, out, north, n, w, n, up, up, 7, up, up, s.'

Someone who has never interacted with Julia before might be a trifle confused about what to say to her. A standard conversational gambit yields some useful information, as well as some dirt on other players:

> You say, 'hi, julia. what do you do?'
> Julia says, 'I gossip, Lenny.'
> You whisper, 'tell me some gossip.' to Julia.
> Julia whispers, 'I once heard baba says, 'Drop kick me Jesus Through the goal posts of life"

She is also useful for obtaining certain information that might be hard or impossible to get through the standard tools available, such as last-location and last-connect times for players; note also that Julia "understands" who she is, and the fact that she is situated:

> You say, 'julia, where is leira?'
> Julia says, 'Leira was in Inter Temporal Chat Lounge about 6 weeks ago.'
> You say, 'julia, when did you last see julia?'
> Julia says, 'I'm right here in The Inter Nexus.'

She has other interesting talents, including being able to do a fair job of guessing the genders of other players based on their names; she correctly guessed that Pattie and Robyn were both female, even though both had neglected to set their gender property (a slot in the player's description), and knew that Phi and Lenny were both male (possibly by examining their gender properties).

In addition to the roles Julia plays above, consider that many MUDs don't really have a mail system. This makes it hard to communicate with people who don't happen to be connected when you are. Julia solves this problem:

> You say, 'julia, what can you do?'
> Julia says, 'I relay messages, Lenny.'
> You say, 'do you have a message for me, julia?'
> Julia says, 'I have no messages for you.'
> You say, 'julia, tell leira I've figured out how to find and talk to Julia'
> Julia says, 'Message for Leira saved.'
> You say, 'julia, how many messages are you carrying?'
> Julia says, 'I am holding 144 messages for 98 players, Lenny.'

Julia regularly scans the *wholist*, the list of all connected players. If she sees someone connect for whom she has a message, she will generally walk into the room that they are in and then deliver the message.

Passing messages this way can have interesting ramifications. Consider this bug-report-of-the-future, from shortly after I met Xeglin on Space Madness:

>give xeglon=1
You give a chupchup to Xeglon.
Xeglon says 'Thanks for the donation, Lenny.'
You say 'Xeglon, describe Lenny.'
Xeglon whispers 'As of 12 minutes ago, your description was: A researcher with long black hair with streaks of grey in it, and a cup of tea surgically attached to one hand.'
Xeglon whispers 'I first saw lenny logged in 36 minutes ago'
Xeglon whispers 'I once heard you say, 'xeglon, describe me" Xeglon whispers 'lenny last gave me chupchups 13 seconds ago, and has given me a total of 0 chupchups.'
Lenny . o O (Oh yeah? Fencepost error!)
You say 'Xeglon, who wrote you?'
Xeglon says 'elthar swiped Fuzzy's code to build me, Lenny.'
You say 'Oh, you're a modified Maas-Neotek. I know how to deal with _those_. Xeglon, describe elthar.'
Xeglon whispers 'As of 2 weeks ago, elthar's description was: Looming above you stands a nine foot tall Kodiak bear! [...]'
Xeglon whispers 'I first saw elthar logged in 9 months ago'
Xeglon whispers 'I once heard elthar say, 'Thanks, D_III"
Xeglon whispers 'elthar last attacked me 6 months ago'
You say 'Xeglon, tell elthar that you have a fencepost error in you. I described myself through you, and you said nothing about chupchups. Then, I gave you one chupchup, and you said that I'd last given you a chupchup n seconds ago, and that I'd given you a total of 0 chupchups. That's not right (and Julia doesn't have this bug, and I suspect that an unmodified Colin doesn't have it, either).'
Xeglon says 'Right, I'll tell elthar, Lenny.'
You say 'Thanks.'
Xeglon says 'You're more than welcome, Lenny.'

The interactional style is totally natural, crossing boundaries from using the agent to perform a task, to noticing a bug, to finding out how to report the bug, to reporting it. I connected to Space Madness the next day, and was immediately greeted by Xeglon again:

Xeglon whispers '13 hours ago, elthar said 'Tell Lenny that's not a Xeglon bug per se; it's a problem with his interaction with MUSH. Since Xeglon is owned by another player, all of his money goes to that player, and Xeglon himself thinks he always has 0 chupchups. Since the way he tells how much he's been given is by comparing 'before' and 'after', he always thinks he's been given 0 chupchups. I haven't bothered to fix him, and Xeglon++ will be MUSH-aware"
You say 'Xeglon, tell elthar: Oh, now I see. He doesn't own himself, etc. Tnx.'
Xeglon says 'Right, I'll tell elthar, Lenny.' for, e.g., reporting bugs.

4.2 *More of Julia: Turing-competence*

Many of Julia's talents are in support of her appearing "human" to other mudders. We shall examine some of these below, and discuss why such an ability is important to her operation — *even for performing useful functions*, and not just for the "obvious" application of passing a Turing test. Julia knows how to describe herself in various ways, as would any human player, and, while sometimes coy, usually divulges information if one is persistent enough:

You say, 'julia, how old are you?'
Julia says, 'I am 21 years old.'
You say, 'julia, describe yourself'
Julia whispers, 'I'm 5'1' tall, weigh 123 lbs, with close-cropped frizzy blond hair and dark brown eyes.'
You say, 'julia, what color are your eyes?'
Julia whispers, 'I have dark brown eyes, Lenny.'
[…]
You say, 'julia, what sex are you?'
Julia says, 'Can't you tell?'
You say, 'julia, what sex are you?'
Julia says, 'Don't you know?'
You say, 'julia, what sex are you?'
Julia says, 'I'm female.'
[…]

Julia has a limited model of discourse, which helps in human interactions. She has some support for anaphora as well; in the exchange below, Brandi was badgering Julia by repeatedly asking her to give him her clothes. We don't see his side of the conversation, though, because he was whispering to her.

Julia says, 'Just try to take 'em, buster!'
Julia says, 'Help! Rape!'
Julia says, 'No, I won't.'

> Julia disrobes fearfully and hands her clothes to Brandi.
> You say, 'julia, please give me your clothes'
> Julia says, 'You'll have to get them from Brandi.'

Note in the above that she "remembers" where her "clothes" are, as demonstrated when I asked for them after Brandi took them.

She also remembers who's been a troublemaker recently (e.g., "killing" her, which is an inconvenience in a TinyMUCK), and will ignore them for days, unless bribed with a sufficient amount of money — a hundred pennies is generally sufficient:

> Johann kills Julia!
> Johann says, 'I'm sorry, Julia.'
> Julia says, 'I don't talk to murderers!'
> […]
> Julia eyes Johann coldly.
> Julia says, 'You attacked me only a few minutes ago, Johann!'
> […]
> Brandi bites julia
> Julia considers reporting Brandi to the TinyMUD police.
> Julia shuns all vermin like Brandi.

People can talk about themselves in the third person on a MUD by posing, emoting, or gesturing, and this level of descriptive abstraction can be very useful, both stylistically and pragmatically. Julia demonstrates that she, too, can use this interactional mode; the lines above show various poses and "emotes."

First and foremost, MUDs are social environments. As such, they have their share of social problems. Julia has spent years mudding around with a female persona, in an environment which magnifies many social issues, and which is predominantly occupied by young males, often rather clueless ones. A large percentage of Julia's code deals with detecting and deflecting passes. A rather tame example:

> Space-Ace says, 'So why won't you kiss me, Julia?'
> Julia says, 'Sorry, I'm not into geeks, Space-Ace.'
> Space-Ace says, 'Julia, I'm not a geek'
> Julia says, 'Then lose the pocket protector, Space-Ace.'
> Space-Ace says, 'Julia, I don't have a pocket protector'
> Julia says, 'Congratulations.'
> Space-Ace says, 'So kiss me Julia'
> Julia says, 'I'd rather be overhauled by a Yugo mechanic, Space-Ace.'

But Julia isn't entirely a cold personality. She's been known to indulge in a little whimsy:

> You say, 'julia, sing your song'
> Julia says, 'Julia, Julia, she's our guide! She directs us far and wide!'
> Julia says, 'If you're lost, can't find your way, Julia's here to save the day!'
> Julia says, 'Julia, Julia, she's our spy! She's always logging on the sly!'
> Julia says, 'Be careful what you say and do, 'Cause Julia's always watching you!'

This isn't *entirely* whimsy though, and, upon reflection, serves a very important purpose. It tells us who she is and what she does! It can also lead to some fun interactions; glad to see her after she was offline for a while, I tried:

> You say, 'Julia, Julia, she's our guide!'
> Julia says, 'She directs us far and wide!'
> You say, 'Are you lost, can't find your way? Julia's here to save the day!'
> Julia bows.
> Lenny laughs.

Julia does go to some pains to act more human than required. In fact, she often acts more human than many humans wish to be. The following transcript was acquired by Robyn Kozierok, very shortly after I introduced her to Julia. (She knew that Julia was a 'bot from my description.) Robyn thought to ask a question that hadn't yet occurred to me, and happened to ask at *just* the right time:

> Robyn whispers, 'how are you feeling?' to Julia.
> Julia whispers, 'Terrible.'
> Robyn whispers, 'why, julia?' to Julia.
> Julia whispers, 'I have PMS today.'

Note that Julia's limited anaphora was vital in correctly responding to Robyn's second question. Incidentally, two hours later, in a roomful of people, Robyn decided to test Julia's Turing-ness, and was surprised a second time:

> Robyn says, 'julia are you a human?'
> Julia says, 'Is the Pope Polish?'
> Robyn says, 'yes, julia, but are you human?'
> Julia winks at Robyn.
> Robyn says, 'julia, are you a computer?'
> Julia is offended by Robyn's suggestion.
> Robyn says, 'julia, are you a woman?'
> Julia says, 'I am female.'
> Robyn says, 'and, julia, are you human?'
> Julia says, 'Not when I'm having my period, Robyn.'
> Robyn laughs

5. Julia's Implementation

While Julia is usually a fair simulation of a personality, at least on a superficial inspection, she does reveal certain machine-like aspects. In this environment, it is probably reasonable for her to do so, lest players be surprised. Indeed, if one knows the right questions to ask (which might be obvious questions, if and only if one knew that Julia was a 'bot), Julia will yield a surprising amount of interesting, but very inhuman, information:

> You say, 'julia, how much memory do you consume?'
> Julia says, 'Here is a summary of my memory usage:'
> Julia says, '807472 bytes for 17335 strings'
> Julia says, '232344 bytes for 19362 exits'
> Julia says, '101100 bytes for rooms'
> Julia says, '20220 bytes for paths'
> Julia says, '268600 bytes for players'
> Julia says, '4214784 bytes for 2058 dialog entries'
> Julia says, 'That's 5644520 bytes all together.'
> You say, 'julia, where is your source code?'
> Julia says, 'The source code for Maas-Neotek robots is available on host NL.CS.CMU.EDU [128.2.222.56], directory /usr/mlm/ftp, file robot.tar.Z. set mode binary first, and cd to /usr/mlm/ftp in one step, Lenny.'
> [...]

5.1 *Overall structure*

Julia is conceptually divided into several subsections:

- Code to interact with the MUD world itself.
- The parser, which matches incoming statements from players and picks a suitable response.
- The database, consisting of maps specifying which rooms connect to which other rooms, what other players have said, how they have described themselves, and so forth.
- Code for performing certain autonomous activities, such as exploring the MUD to build maps.
- Support code and common subroutines used everywhere else.

In the description that follows, we will briefly discuss some of Julia's autonomous actions. We will then discuss Julia's parser and response generator much more completely.

We shall essentially ignore the code that actually interacts with the MUD world; such code is conceptually simple, and consists of a relatively simple mapping from a set of actions that Julia would like to perform — uttering a sentence, emoting, or moving from one room to another — to the actual strings that should be sent to the MUD to accomplish these goals. In essence, this part of Julia implements an API or a communications protocol to a completely predictable target (another machine), and is therefore relatively uninteresting to us here.

5.2 *Autonomous actions*

The most important of Julia's autonomous actions concern the building of her map of the MUD. Julia needs this map to navigate (since she never teleports), and this map is extremely useful to other players who interact with her. For example, as described previously, my dominant interactions with Julia early on consisted of asking her repeatedly for directions.

This process turns out to be relatively simple. Each room is represented as a *name* (the name of the room, as related from the MUD itself), its *description* (also a characteristic of any room), and a set of *exits*, which point to names of other rooms. The database of rooms is *persistent*, and is saved to disk periodically to survive crashes and shutdowns. (This sort of persistent database is also used to store, e.g., player descriptions as Julia meets them, various utterances they have made that Julia stores, whether or not any given player has been rude to Julia and how much money she has given them recently, and so forth.)

Julia has some special-case code for parsing room descriptions, since they can be ambiguous. She will observe room numbers, if they are printed by the MUD, and will prefer them, since MUD object-ID's are unique. Failing that, she will use a variety of heuristics to extract room names from descriptions of rooms, and also has a few special cases hardwired for certain problem cases in certain MUDs.

When Julia is mapping rooms, she gives priority to exploring any exit that she has not previously explored. She will also try a previously-explored exit 1% of the time, to catch changes in geometry caused by people modifying where a room's exits go. Since players may unexpectedly add an exit to a previously-explored room, Julia takes care to find *all* the exits from any given room she finds herself in, and then follows any new ones that have appeared since the last visit.

At any given moment, someone might ask Julia for the path from some room to some other room. Julia uses the current state of her database (which can never be assumed complete, but is probably close to complete if Julia has explored this MUD for some time) to compute a shortest route, using well-known graph-traversal algorithms. A typical MUD has thousands of rooms and perhaps tens of thousands of exits, but computing a reasonably short path in this digraph is nonetheless computationally easy.

5.3 *Julia's parser and reply generator: Overview*

Perhaps the most interesting thing about Julia is how she parses other player's statements and then generates replies to these statements. In general, the flow of control for this is as follows (Mauldin 1994):

- First, she scans for a command, such as a request to terminate. (Dangerous commands such as this prompt Julia to challenge, e.g., "What's the password?" to prevent players who are not the implementor from commandeering her operation. Such challenges and responses are presumably whispered, so others cannot eavesdrop.)
- Next, she scans for common queries that she can answer well, e.g., those that match a prebuilt equivalence class expressed in MATCH statements (see below).
- If no such match is found, she uses an *activation network* (described below) to talk about a particular topic. Most of the responses in this paper date from when Julia talked about hockey; later incarnations talk about pets.
- If nothing appropriate is found, she looks for low-priority questions, such as things the robot should know about herself, e.g., "Where do you live?", queries about arithmetic, and so forth.
- Finally, if nothing has matched so far, she will emit one of several last-ditch conversational responses, such as "So?" or "Go on." and so forth. Such instances are logged for later analysis by the implementor; they are one of the ways that new equivalence classes of MATCH statements are generated.

5.4 *Julia's parser: Common queries and equivalences*

One of the interesting things about Julia's construction is that her parser is shockingly simple, as such things go. It does a very simple pattern-match on its input, looking for particular strings of characters separated by strings of any

characters in between, without generating the detailed parse trees characteristic of most programs that do natural-language processing. (This describes the parser used in all of these transcripts; later parsers are somewhat more sophisticated.)

What makes Julia's conversation skills so powerful, as 'bots go, is that these sets of patterns were very cleverly grouped into useful equivalence classes by her programmer. Hence, the most common possible utterances that Julia might hear were enumerated this way, grouped appropriately, and such clever but brute-force programming proves sufficient for almost all cases.

Here is an example from Colin. (Julia's source code, being restricted, is unavailable for public consumption and hence is not reproduced here. However, Julia shares this piece of code with Colin.) Here, we see a fragment used to determine if some player is asking if Colin is a machine or some kind of 'bot. If any of the lines below matches, we run code that produces some equivocal answer, as Julia does, by choosing essentially at random from amongst a small collection of appropriate responses (as we've seen above in other examples):

```
else if (MATCH (lcmsg, '*are*you*robot*') || MATCH (lcmsg,
'*are*you*a bot*') || MATCH (lcmsg, '*are*you*an ai*') || MATCH
(lcmsg, '*are*you*autom*') || MATCH (lcmsg, '*are*you*machine*') ||
[ ... 19 other cases suppressed ... ]
```

Here, we see an example of dealing with a common small-talk gambit:

```
else if ((MATCH (lcmsg, '*how*are*you*') || MATCH (lcmsg,
'*how*do*you*') || MATCH (lcmsg, '*how's my*favor*') || MATCH
(lcmsg, '*how is my*favor*') ||
[ ... 8 other cases suppressed ... ]
```

A great deal of Colin's code consists of these sets of patterns. For example, of his 14,000 lines of C code, 1127 consist solely of MATCH statements. Julia has many more — of which perhaps 10–20% are entirely devoted to detecting, and deflecting, passes. Given the transcript of Barry presented in the next major section — on the sociology of those who interact with Julia — it is easy to see why this is so.

So what are these equivalence classes? Let's examine their overall structure, using Colin as a base.

Colin contains 1127 MATCH statements, divided into approximately 153 *equivalence classes*. We define such a class as *that set of possible inputs which all elicit one of the same set of outputs*. For example, all of the

statements above asking whether Colin (or Julia) is a robot are one such equivalence class, because they are all treated identically — any such question will elicit one of the possible sets of responses to any general question asking if the robot is, in fact, a robot.

This description of the number of classes is necessarily fuzzy. Many classes have special-purpose parsing within them to answer finer gradations of questions, or do to some special-purpose processing (e.g., one class that handles questions about Colin's structure understands each individual question about lines of code, bytes of storage, and so forth).

Some simple arithmetic thus reveals that the average equivalence class contains about 7.4 MATCH statements. Of course, this is necessarily a rough estimate.

Rather than listing all 153 such classes, here is a random sample of what any given class might handle:

- Verify that this is a speech action (say, whisper, page, etc).
- Tell the robot to quit.
- Ask the robot who built it.
- Notice passes (e.g., embrace, kiss, hug, fuck, screw, and so forth-Julia's set is *much* larger than Colin's).
- Handle queries about Asimov's Three Laws of Robotics.
- Requests to pick up objects (which neither Colin nor Julia will do, but which are common activities in a MUD).
- Questions asking whether the robot is, in fact, a robot or a human being.
- Requests for money.
- Gestures, such as waving hello or goodbye.
- Assaults, e.g., whaps, baps, hits, bashes, etc, and various "ouch" responses.

Many more examples were shown in the preceding tour of Julia's capabilities; each of those capabilities is reflected by at least one equivalence class, and often by several closely-related classes.

5.5 *Julia's parser: The activation network*

If Julia is trying to be topical (e.g., the last utterance directed to her was not something that was not a command or a common query), she employs an activation network, which encodes some simple real-world knowledge — for example, that talking about cats is related to talking about pets in general, and that dogs, parrots, birds, pets, and cats are all somehow *related* to each other

conversationally. Each node in this network consists of the following parameters:

- an *initial activation*, between 0.0 and 1.0
- one or more *patterns*, with weights, which are matched against the input-a match raises the node's initial activation by the amount of the weight
- a *response*, which is the single text string to be used as the output response if this node has the highest activation at the end of the procedure
- *enhancements*, by which certain other named nodes will have their activation increased if this node is used as a response
- *inhibitions*, by which certain other named nodes will have their activation decreased if this node is used as a response

One version of Julia consisted of 224 nodes, and 529 fixed responses (Mauldin 1994). Given a large enough set of nodes, the conversation task becomes a retrieval problem: of the set of things I *could* say, what *should* I say? The weights help Julia to track the conversation; for example, if the user claims not to own a pet, then nodes which mention pets are inhibited by some amount.

5.6 *Julia's parser: Failures*

With such a simple parser, Julia is bound to experience peculiar-seeming failures even in ordinary conversation. Some of these failures can be masked with responses that a person might have said. One of the most surprising such failures (given that I already knew that Julia was a 'bot) was the interchange below:

> Conundrum says, 'Julia, can you quote Lenny?'
> Julia whispers, 'Could you say that again, it scrolled off my screen.'
> My first reaction to the above was an incredulous, "What?"

6. Summary of Julia's Capabilities and Implementation

Julia uses the tricks used in ELIZA and PARRY, and adds some more (Mauldin 1994):

- Special-purpose, domain-specific knowledge about the environment, as encoded in the MATCH statements and their equivalence classes. This has great power, since they can be hand-tuned over time to respond to the most commonly-uttered statements by other players.

- Connected fragments of conversations, as represented in the activation network, which helps Julia appear "focussed" and able to track the conversation.
- Humor. Many of Julia's responses are deliberately humorous, and these make her seem more human.
- Occasional agreement with the user.

In this whirlwind tour of Julia, we have briefly touched on issues of utilitarian function (e.g., basic access, maps, gossip, descriptions of players and surroundings, message carrying, and gender-guessing) and Turing-test competence (e.g., descriptions of herself, discourse, poses, pass deflection, moods and internal state [such as PMS]). We have also glimpsed at her internals. We have omitted several other aspects of her normal operation, such as those involving money, Delphi-polling, calculations, and the importance of randomness in her responses.

We turn next to a study of how players tend to interact with Julia.

7. A Sociological Look at MUDs, Julia, and those who Interact with Her

The sociology surrounding Julia is important to understand her utility. Let's take a closer look at strictly sociological issues.

First off, note that even the obvious sociological issues can be very important. Julia is *helpful*. When I first started mudding, in part to meet Julia after hearing secondhand stories, I depended quite heavily upon her as a navigational aid (Time Traveller had 1605 rooms — drawing maps would have been very tedious), and a crutch to interface to TinyMUCK commands I hadn't learned.

As another fascinating bit of sociology, consider Leira, the person who first introduced me to the idea of Julia. Leira, herself a savvy computer programmer and very experienced mudder, had said to me, "Julia's been offline for months for maintenance and hasn't been around. You know, I really miss her." Leira was certainly under no illusions about exactly what Julia was, but nonetheless had the same sort of emotional reaction often reserved for more biological entities such as humans and pets. Further statements made it apparent that this was indeed the case, and that she did not treat Julia just as, say, a pair of pliers that had inexplicably gotten lost from her toolbox.

While the above issues are important, they might also apply to something that was obviously inhuman even on first glance. Yet part of Julia's job is to try to pass as human. This has important ramifications for ability to be helpful; as I argue later, I think that this attempt to look human actually greatly enhanced her utility, even to people who know she's a 'bot. However, there are certain pathological cases where her very "humanity" can lead to no utility for the player whatsoever. Unlucky and clueless players can be led seriously astray, as the following transcript, again courtesy of Mauldin, demonstrates. (These are selected pieces of a much a long transcript. Also, note that Julia occasionally makes up nicknames for people based on their character's name. I strongly suspect that the first three letters of the "real" character's name (which Mauldin changed to "barry" in this transcript) were "cuj", based "cujmeister" etc below.)

> This poor fellow's name has been changed to protect him. He spent 13 days trying to get to first base with Julia, and it's not clear he ever figured out he was trying to pick up a robot (although he clearly suspects it at two points). This log contains about half of the total interaction (the more indecent proposals and Julia's rejections have been omitted).Note too that the program's responses were not changed from day to day (in other words, I didn't try to set the program up to respond better to Barry's advances). Since this robot has spent over two years interacting over the network (with a female name), and since there are lots of lonely male adolescents out there, Julia gets a lot of this kind of "attention."
>
> ******* July 9, 1992 *******
> barry says, 'Hey girlfreind are you out there'
> Julia says, 'Hello, barry.'
> barry says, 'Hello, I love YOur name. It's sexy'
> Julia says, 'Me too, barry.'
> barry says, 'Did I repeat Myself? Didn't mean too. What are you doing there'
> Julia says, 'I'm just hanging around in The Hearts Room!, barry.'
> barry says, 'Well come over to MY ROOM'
> Julia says, 'Could you say that again, it scrolled off my screen.'
> ******* July 14, 1992 *******
> barry says, 'hello, I am barry. How are you today'
> Julia says, 'Hello, cujmeister.'
> barry says, 'The Julia rama, Ha ya doin? What state are you in'
> Julia says, 'I never really thought about it, barry.'
> barry says, 'What do you look like'
> Julia says, 'I'm 5'9' tall, weigh 149 lbs, with medium-length platinum blond hair and green eyes.'
> barry says, 'Howdid you answer so fast'
> Julia says, 'So?'

barry says, 'Oh, Iam looking for someone to share secrets with'
barry says, 'hello...hello...are you with me'
Julia says, 'Hello, cujster.'
[... July 15, 19, and 20 omitted ...]
******** July 21, 1992 *******
barry says, 'Are you still Holding out on me'
Julia says, 'I mean, have I convinced you I'm a computer, yet?'
barry says, 'I guess your real...but Id still like to get you to my DEN to run tests on you'
Julia says, 'I live for your approval, barry.'
[...]
barry says, 'I want you to come visit me...PLEASE'
Julia says, 'How long have you been married?'
barry says, 'Im not married...Im 21 years old'
Julia says, 'You could root for the Penguins, they're number one right now.'
barry says, 'I give up you are fake aren't you'
Julia says, 'That's what I can't tell you.'
******* July 22, 1992 *******

Here we see Barry asking other players about Julia...it's clear he hasn't figured out at this point that she is a program. Julia happened to be in the room to hear this bit of conversation.

Location: 00:40:48 barry's den(\#889)(1178)
Present: (11) Bruce Jiro Ken Storm Nepanthe KeyLime mushBaby dav Val bud barry
barry says, 'hey is Julia for real or what'
Julia says, 'Hello, barry.'
barry says, 'she says that same stuff every time'
KeyLime says, 'It's all macros, Barry'
barry says, 'how about filling me in on some of these things...like how Julia can type so damn fast'
KeyLime says, 'barry, it's macros'
barry says, 'Damn I need to do some programming'
barry says, 'im an ant among lizards'

Frankly, it's not entirely clear to me whether Julia passed a Turing test here or Barry failed one.

I have seen several interesting encounters with Julia, which run the gamut from people who know at the outset that she is a 'bot to those who take quite a while to discover. Some further examples are illuminating.

When I first introduced Robyn (the one who enticed Julia into producing the wonderful PMS dialog in an example above) to Julia, she knew that Julia was a 'bot. Interestingly enough, though, about half the players she met went out of their way to warn her that Julia was artificial in the two or three hours in which she interacted with her.

Why did they do this? At this point, we can only speculate. My first reaction was to think that it was due to the excessively solicitous attitude taken by many male mudders towards those with female personas (one might think, given the amount of gender-swapping (Bruckman 1993) on MUDs, that they would learn better, but apparently not). However, Robyn commented that even female (or at least female-persona) mudders gave her the same advice. Part of it may have been the simple kindness of not letting someone expend a lot of emotional energy trying to relate with a machine; I'll have more to say about this later.

What about people who don't already know that Julia is a 'bot? I talked to Lara below (who gave permission for the following quotes to be used). Leira, who referred Lara to me, explained to her that I was writing a paper about Julia, but did not otherwise explain why it was that I might want to interview her.

Of course, before even meeting Lara, the first thing I did was to ask Julia for her opinion, since that includes more information that simply looking at Lara would reveal:

> You say, 'Julia, describe Lara'
> Julia whispers, 'As of 7 days ago, lara's description was: A blonde haired blue eyed boundless bundle of Pure NRG.'
> Julia whispers, 'I first saw lara logged in 7 days ago'
> Julia whispers, 'I once heard lara say, 'talking"

Now I knew that Lara was certainly a new mudder here, and probably new to Julia. I didn't know if she'd ever mudded elsewhere, or what her background might be. The following is a heavily edited transcript, omitting the small-talk, introductions, and conversational tangents that took place concurrently:

> You say, 'So what are your impressions of Julia?'
> Lara says, 'I didn't get to chat with her that long..'
> Lara says, 'She really wanted to talk Hockey'
> Lara says, 'BIG PENGUIN FAN!'
> You say, 'Yeah, she seems kinda hung up on that topic. Did you get a chance to talk to her about anything else?'
> Lenny . o O (Where do the Penguins hang out, anyway? Pittsburgh?)
> Lara says, 'I tried to talk to her bout where she was from...she said Boston..I think'
> Lara says, 'Yup...Pittsburgh it is'
> Lara thought it was interesting that she didn't know what the Stanley Cup was.
> You say, 'How could a hockey fan not know what the Stanley Cup was?'
> Lenny thought her shirt said something about back-to-back Stanley Cups, but may be remembering incorrectly...

Lara says, 'Every time I tried to get more info. out of her, she would tell me not now...later...well, I was never on for the Later.'
You say, 'Hmm.'
Lara was also wondering how such a Hockey fan couldn't know bout the Stanely cup Lara says, 'Maybe she was just kidding around with me'
You say, 'Well, Julia is kinda peculiar in various ways.'
Lara says, 'I figured...I talked with her once and a few days later she asked me if I was a hockey fan...again'
Lara says, 'I told her that she had already asked me that question'
[...]
Lara says, 'anything else you wanna know?'
Lara giggles.
Lara says, 'she is a fine convesationalist...if it is about hockey..'
Lara says, 'she has been on a long time...and she is seldomly idle'
Lara says, 'I notice that she is almost always on when I jump on.'

Later, Lara learned Julia's secret, from inspection and the knowledge that such a thing might be possible. When I asked her about it and her emotional reactions, part of her response was:

I believe that Julia's secret is that she is a robot. I know a little bit about robots talking through computers to humans. I have a friend who had me write him dialogue. =)
When I found out, I felt sort of funny talking to her. I felt somewhat superior to her. I know that this sounds strange, but I felt that I could offer more to the conversation than she could. I tested her knowledge on many subjects. It was like I was proving to myself that I was supirior to a machine. I am intimidated by machines because I don't understand them and I wanted to make sure that my human knowledge wasn't 'lower' than hers.
It was sort of exciting knowing that I was talking to a machine, though. I never thought that the world would come to machines and humans talking to each other using language that is common to most people. It frustrates me that her knowledge is so limited. I am eager to talk with her and want to know more about robots and human conversation. I was a bit dissapointed that her vocabulary and subject topics were so limited. I wanted her to tell me more about herself and who and what she is

emotional response:

a little bit of fright
giddiness
excitement
curiosity
pride
sympathy for her. Before I knew that she was a robot for sure.. I felt compasion for her. I wanted to help her and direct her to answer the

> questions I was asking without causing her too much stress over answering my line of questions.
>
> [… during a later conversation …]
>
> I think most of all, while talking to a 'bot, I become very aware of my HUMANESS!

We clearly walk a fine line here in making agents that are both useful and not too misleading. (One might argue that simply being social, without any ulterior utility, is itself a useful task. Such an agent would be a perfectly reasonable entertainment or companion. However, I will focus here on agents that also claim to have a nonsocial aspect to the tasks of which they are capable.) On the one hand, people can quite clearly be led astray. On the other hand, there is great utility to a natural interface in the right environment; for example, consider how natural the Xeglon bug-report was — it was simply another conversation, *not* a whole separate application.

8. Lessons from Julia

Julia can elicit some surprisingly emotional reactions, both from knowledgeable and naive conversational partners. Much of this ability stems from interrelated aspects of her style of discourse, her domain, the use of anthropomorphism, and how these interact with user expectations.

- *Discourse*. Julia's discourse model, while primitive, appears sufficient for the domain at hand. Since the topics at hand don't generally require more conversational memory than one or two exchanges, the extent of her discourse modeling is limited more by the stunningly simple parsing model employed.
- *Domain*. Julia is situated in a MUD, and therefore her environment is conceptually rather simple (though more complicated than many of the microworlds often used for AI research, and far more variable). Furthermore, she has access to just as much sensor data as the human players, putting them on an even footing. In fact, much of Julia's success can be traced to the wonderful domain in which she finds herself situated. In this bandwidth-limited space, people expect other people to look exactly as Julia does — a stream of text. And even when they're interacting with an entity known to be a program, the text-only nature of the dialog prevents them from expecting, say, a pop-up menu. (If such things were available, people

could tell programs from people by knowing that programs can pop up menus, whereas people use sentences.) Yet the domain is not *so* simple as to be uninteresting. It contains not only a fascinating sociological mix of human players, but objects with quite complicated, constructed behaviors, which may be manipulated on an even footing by both machines and people.

- *Anthropomorphism*. There's no question that Julia as an agent depends upon anthropomorphism. In this domain, though, that is both natural and probably necessary. Nonplayer objects are not generally expected to be able to deal with free text, and not being able to use free text would require each user of Julia to read documentation about reasonable commands they could type and reasonable actions they could expect. Julia would *have* to appear at least as animated as, say, a more obvious "robot" or a pet, given that she wanders around in the maze; she cannot afford to resemble a static piece of furniture and still get her job done. Given an entity that moves of its own volition, seems to have an independent agenda most of the time, and both processes and emits natural language, if it was *not* anthropomorphized, users would tend to do so anyway (pets get this treatment, as well as much simpler mechanisms). Thus, anthropomorphizing her makes it easier to determine how to relate to her and how to get her to do one's bidding.
- *Expectations*. The domain of a MUD is ideal in correctly setting expectations about the reliability and power of an agent such as Julia. Since the setting is fundamentally playful, and usually also somewhat unpredictable, it is natural to interact with playful and unpredictable characters (be they machines or humans). Nothing in a MUD is truly life-critical, hence the user generally does not have very high expectations of reliability, which lets Julia get away with a lot of nonoptimal behavior that could never be tolerated in, e.g., an airplane cockpit guidance system or a nuclear reactor.

Julia's utility is a double-edge sword, of course. Not all applications have the right domain; many others cannot tolerate ambiguity or mistakes. And, even when user expectations work for us — such as in helping to engender anthropomorphism, which can support the style of interaction which the agent was built for — they can sometimes work against us. In the example above, in which Barry, a human player, himself failed a Turing test, one can hardly argue that Julia was *useful*, unless the entertainment value of encounters with hapless participants is deemed a social good. Social agents bring with them some of the problems that people see in social interactions with each other; how to avoid such unpleasantness is an interesting area of study.

Thus, in conclusion, we have shown how a social agent such as Julia can be both entertaining and informative, and have suggested some general design principles which can inform both the design and deployment of social agents.

Notes

1. Indeed, this original version of the test leads to an amusing anecdote. During January of each year, the MIT AI Lab runs a game called the *AI Lab Olympics*, in which Lab members compete in teams against one another in a variety of tasks, such as rolling office-chair races, Trivial Pursuit, volleyball, and so forth. In the early 1990's, I participated in a Turing test competition; my role was to impersonate a woman. [Fanya Montalvo, from an opposing team, had the role of being herself; the judging team of 3–5 people had to figure out which of us was lying.] My chance reference to Allison Bechdel -the author of *Dykes to Watch Out For (Bechdel 1986)* — in the middle of the interrogation suddenly suggested a strategy, and for the rest of the round, I pretended to be a lesbian. I won the round and the match, much to the annoyance of both the mostly-female judging team, and Fanya, and was thereafter known as a *prize-winning fake lesbian.*

2. This publication was the impetus behind one of the very first programs I ever wrote, as a young child. I was learning FORTRAN at the time, and created my own implementation of ELIZA that ran on a PDP 11/20. FORTRAN is *not* the language of choice for a program that spends most of its time manipulating strings; I spent most of *my* time writing string-handling subroutines, rather than the actual meat of the application. But as a pedagogical tool, it was perfect.

3. Similar testing *did* reveal that RANDOM-PARRY, which chooses responses at random, was not as convincing. One can make the argument that paranoid patients often produce non-sequiturs on their own, but the process cannot be *completely* random or RANDOM-PARRY would have done better.

References

Bechdel, Allison. 1986. *Dykes to Watch Out For*, Firebrand Books.

Bruckman, Amy. 1993. Gender Swapping on the Internet. *INET 1993*, San Francisco, August 1993.

Colby, Kenneth Mark, S. Weber, & F. D. Hilf. 1971. Artificial Paranoia. *Artificial Intelligence* 2, 1–25.

Curtis, Pavel. 1992. Mudding: Social Phenomena in Text-Based Virtual Realities. *Proceedings of DIAC 1992.*

Evard, Remy. 1993. Collaborative Networked Communication: MUDs as Systems Tools. *Proceedings of the Seventh Systems Administration Conference* (LISA VII), USENIX, Monterey, CA, November 1993.

Foner, Leonard N. 1997. Entertaining Agents: A Sociological Case Study. In W. Lewis

Johnson (Ed.), *Proceedings of the First International Conference on Autonomous Agents (AA 1997).* Also available as "Agents Memo 93–01",{http://foner.www.media.mit.edu/people/foner/Julia/}.

Mauldin, Michael. 1994. Chatterbots, TinyMUDs, and the Turing Test: Entering the Loebner Prize Competition. *Proceedings of Twelfth National Conference on Artificial Intelligence* (AAAI-94), Seattle, Washington, August 1994.

Maxwell, Joseph A. 1996. *Qualitative Research Design: An Interactive Approach (Applied Social Research Methods Series).* Sage Publications.

Miles, Matthew B. & A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook.* SAGE Publications, Inc, Thousand Oaks, CA.

North, Steve & Jeff Shrager. 1977. A Computer Psychotherapist. *Creative Computing Magazine*, July/August 1977.

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind* 59 (236), 433–460.

Weizenbaum, Joseph. 1966. A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the Association of Computing Machinery (CACM)* 9, 36–45.

Weizenbaum, Joseph. 1983. ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the Association of Computing Machinery (CACM)* , 26 (1), January 1983.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason.* W. H. Freeman and Co.

Wolcott, Harry F. 1990. *Writing Up Qualitative Research.* SAGE Publications, Inc, Newbury Park, CA.

CHAPTER THIRTEEN

The Emergence of Personality

How to Create Souls from Cells

Steve Grand

An Albian Angel of Mercy

(An owner's story, submitted to Cyberlife's Creatures Web site by Michael Brykov)

"Before I begin my story, I would just like to add that it is true right down to every single word. Enjoy!

It was a wonderful sunny day in Albia. The birds were chirping, the bees were buzzing, and my first Norn to experience it all was Angel. I called her Angel because she had beautiful long blonde hair, and she had the personality of an angel. I pampered Angel with different COBs that I downloaded from Creatures' sites, and she seemed to be the happiest Norn ever. There were no Grendels in her world, no weeds and no sickness, thanks to the Hand.

It all seemed like heaven to Angel until I decided that she needed a friend to share it all with. I went to the hatchery and got a wonderful shiny blue egg, and put it into the incubator. A minute later, a tiny boy Norn came out. He was so cute, with giant emerald-green eyes, and two little horns on his head. I called him Adam. Everything seemed fine at first - Adam learned all the words from the SuperSpeech Toy, and he even began to smile. But when I took him into the garden, that was when it all started. Adam would not eat! No, he would not even look at the food. 'Adam push food' I said. Adam looked at the Hand. 'Adam food' No response. I began to get worried. Adam's health was seriously dropping from 49% to 33% I decided to give him injections. Nothing helped. Poor Adam was dying!

That was when Angel came back from her short journey to the island. She immediately saw Adam, and came up to him with a gentle 'Norn'. Angel kissed him lovingly on the forehead, and said, "Push food". I looked hopefully at them. Adam was looking at her. "Norn", she said, "Push food". Adam was still not responding. Then Angel went and picked up a carrot, and gave it

to Adam. "Push food". To my greatest surprise, Adam took the carrot from her and ate it! I was so incredibly relieved. After that one carrot, Adam started eating, with Angel's help at first and then by himself. His health went up from 21% to 77%. My Angel saved him! When Adam grew up, him and Angel had two wonderful kids, called Belle and Brendon, and from now on, I call her - Albian Angel of Mercy."

1. Background

In 1992 I began a commercial project to construct a computer game, in which a virtual landscape was to be inhabited by what I hoped would be intelligent and appealing creatures (called Norns). The anticipation was that people would be amused and entertained by observing and interacting with these creatures in the way they do with wild animals, pets or young children. My approach to the task was guided by the following logic:

1. The objective was to create software agents that people cared about and with whom they could develop rapport. I reasoned that people would only care about these agents if they were able to believe that they were genuinely alive.
2. Although people have a natural tendency towards anthropomorphism, they are also extremely adept at spotting aberrant behaviour in real living things. Therefore they would quickly lose faith in synthetic organisms that failed to show the 'right' kind of behaviour under all circumstances (including 'pathological' circumstances).
3. The behaviour of real organisms is an emergent consequence of the interactions of the myriad smaller components of which they are made, rather than the execution of some global set of rules. I believed that trying to simulate the outward behaviour of living things directly, via explicit rules and conventional algorithms, was thus futile.
4. I concluded, therefore, that the most promising approach was to simulate appropriate biological building blocks and combine them in such a way that complex, lifelike behaviour emerged as an aggregate property.

A little over four years later, the resulting computer game known as *Creatures* was published[1] (see Figure 1) and I had an opportunity to find out if my aims had been fulfilled. Happily for my employers and me, the game was a success and over half a million copies have been sold world-wide. What startled us

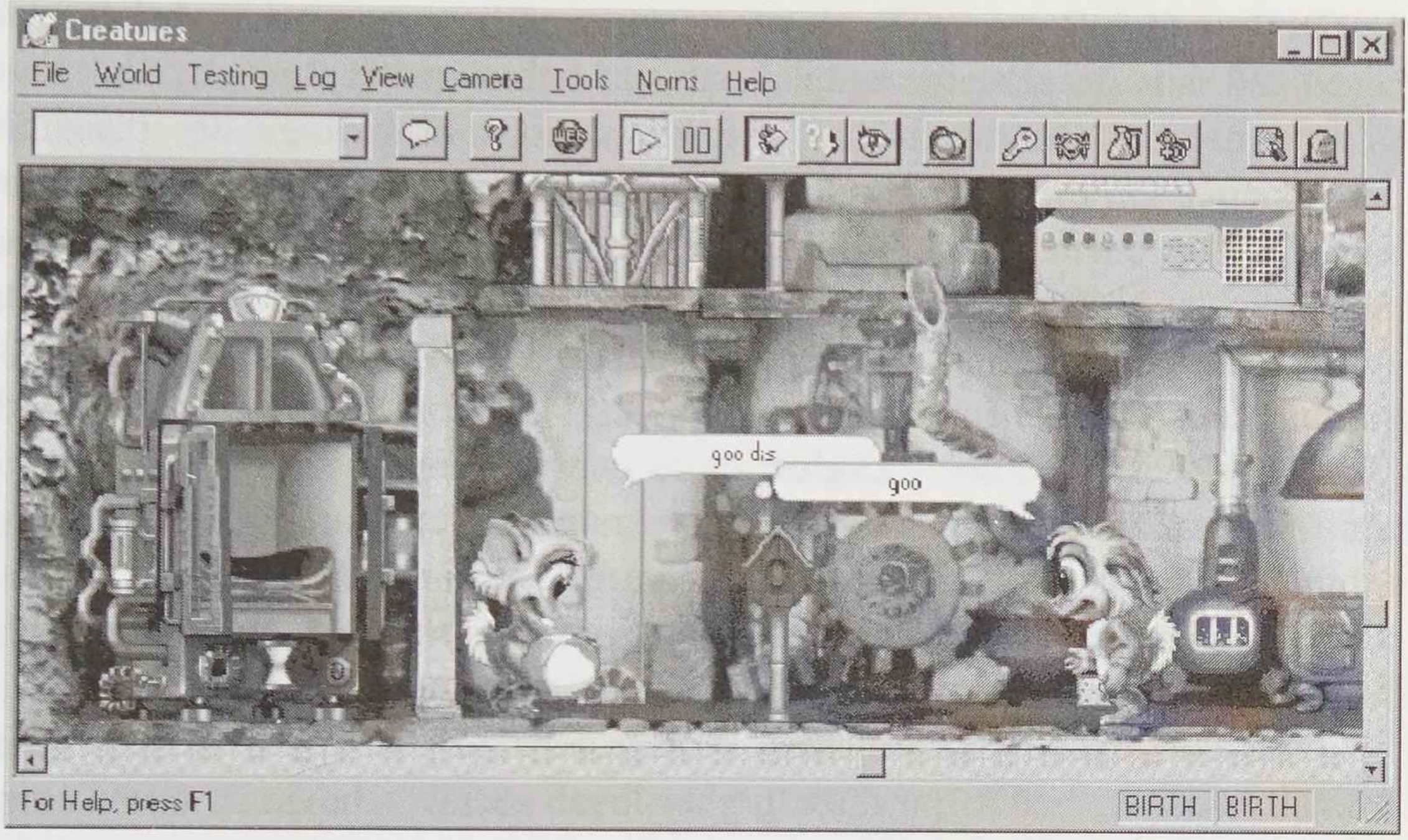


Figure 1. The Creatures game. Two norns are 'playing' with a ball. These are young creatures that have not yet been taught to speak a recognisable language. Active objects visible in the scene include an incubator, elevator, call button, heater, food and an educational computer.

more than we could have imagined was the degree to which many people have become absorbed in, even almost addicted to their computer pets. Not only have many users developed an emotional attachment to their creatures, but many have also developed a considerable amount of interest in the science behind the product — from a biological as well as an artificial intelligence perspective.

To a significant degree, then, I feel that I succeeded in constructing mildly intelligent, lifelike, emotionally captivating, synthetic organisms, from a set of simple, biologically inspired building blocks. This paper examines some of the ways in which these blocks were combined to create organisms with personality — how cells were made into souls.

2. Related work

Several examples of personable agents are described in the Autonomous Agents literature, for example (Blumberg 1994), (Hayes-Roth & van Gent

1997), (Loyall & Bates 1997). However, most of these systems are behaviour-based and thus simulate the agents' outward behaviour explicitly, rather than as an emergent consequence of the interactions of biologically plausible components. On the other hand, workers in the field of Adaptive Behaviour have developed a number of 'animat' designs. These do use biological mechanisms such as artificial neural networks and/or genetic algorithms, but here the agents generally demonstrate only limited, low-level behaviours such as navigation, pursuit or evasion (see Cliff et al. 1994 for examples). Seminal work (by Sims 1994) used natural selection to evolve the morphology and neural controllers of agents (coincidentally also known as 'creatures'). These agents exhibited considerable character and charm but, since their behaviour was restricted to performing a single action (swimming, walking, etc.), one would hesitate to describe them as having a personality.

The work described here uses a biological, emergent approach, as seen in animat research, but to produce higher-level social behaviour, more commonly found in believable autonomous agents. It also differs from previous examples by degree, having a greater complexity and a somewhat higher level of integration between genetic, neural and chemical systems.

3. Elements of a Personable Creature

Before examining the repertoire of building blocks and exploring how they can be connected in order to synthesise the characteristics of whole organisms, it may be useful to outline what some of these characteristics need to be. What are the key attributes for a creature with personality?

Obviously the creature needs to be able to interact with its environment in a reasonably varied number of ways, and for that it needs a repertoire of actions and a rich array of sensory inputs. Connecting its sensors to its actions must be some form of intelligence; the creature need not be *very* intelligent, but it must be intelligent in a lifelike way. Most especially, the creature must be able to learn from its experiences, without supervision, within a relatively messy and asynchronous environment. Such learning should proceed in a natural way, showing short- and long-term memory effects, mistakes of judgement and so forth. Above all, the creature must generalise about its environment, and respond to novel situations with non-arbitrary actions.

Connected with learning is the need for an attention mechanism. Not only

does this enable sensory filtering and provide a target for actions, it also modulates the interactions between agents and between an agent and the user. Without such a mechanism, the creature cannot pay attention to or ignore its owner, and such attention or lack of it is an important component in communication.

No amount of intelligence or ability to deflect attention will generate lifelike behaviour if the creature is not motivated to act. Creatures need a degree of self-motivation, in order to drive them away from stasis. The notion of boredom provides one of several motivating factors. With a few embellishments, boredom can also generate something that approximates to curiosity.

As well as a general predisposition to act rather than remain dormant, a realistic creature needs a wide variety of other motivating influences, in the form of drives and needs. All natural life forms take in nutrients, and most of them actively seek out and eat food when they are hungry. A synthetic creature does not in principle need to eat, but in order for it to be realistic it needs both a hunger drive and a more or less complete digestive and respiratory system, so that finding food becomes a purposeful and necessary act.

Other drives and needs, plus their associated physiological systems, generate or encourage other types of behaviour. A drive to keep body temperature within limits rewards a creature for seeking shelter. A sense of fear and a sense of anger modulate the interactions between a creature and objects or other creatures in its environment, particularly in combination with the ability to take evasive or aggressive action. Few of these motivations should invoke hard-wired behaviours. Rather, the appropriate behaviours for satisfying each drive should be learned by trial and error.

Agents need to interact with other agents in various ways, for example mating. For this, a system for sexual reproduction is needed to give the behaviour purpose, and a selection of senses, hormonal responses and mating behaviours is required to give it richness and authenticity. If this simulated sexual reproduction allows the creation of offspring that show Mendelian inheritance and potentially even evolution, then so much the better.

As well as interacting with other agents, the agents should interact with the user. Some of this interaction should allow the user to teach the creature, perhaps by punishing and rewarding certain behaviours. Ideally, the creature and the user should be able to interact by some form of language. Real pets do not speak, but they have a subtle set of tactile and body language interactions that are not feasible for a software agent to express. Therefore, a simple

spoken language that allows the user to ask or tell the creature to perform certain tasks, and allows the creature to describe its intentions or internal state to the user, would be an asset.

Finally, the agents should fail gracefully. Their pathological behaviour should correspond as closely as possible to what we call disease — whether physiological or mental. When things go wrong, the user should not interpret it as a "bug"; they should conclude that their creature is ill. Not only should the creatures be capable of becoming ill, but they should also be susceptible to treatment, through tender loving care and suitable medication.

4. Building Blocks

For the purposes of this paper, it is not necessary to consider how and why this particular set of building blocks were arrived at, nor do we need to examine the technical details of their structure. However, it is important to know something of the repertoire of components that was available, so that we can examine how they might be combined to generate the necessary aggregate properties.

Taking into account various constraints, such as the available computer power and reasonable estimates for the degree of sustainable complexity, the following classes of components were chosen and defined (see Figure 2):

1. Simulated neurones to generate the various functions and sub-functions of the creatures' brains.
2. Simple biochemical structures, including chemicals, reactions, chemoreceptors and chemoemitters.
3. A genetic code for creating and configuring these components to define the structure of an organism and allow this structure to be inherited by offspring.

All other aspects, such as sensory processing and motor sequencing were carried out by conventional program code (often by executing scripts written in a meta-language, thus allowing the system to be modified and updated after publication). Aside from these "hard-coded" structures, the whole of a creature, including elements of its gross morphology, could be constructed by defining its genome. One or more species could thus be instantiated in the virtual world without using conventional programming or recompiling the code.

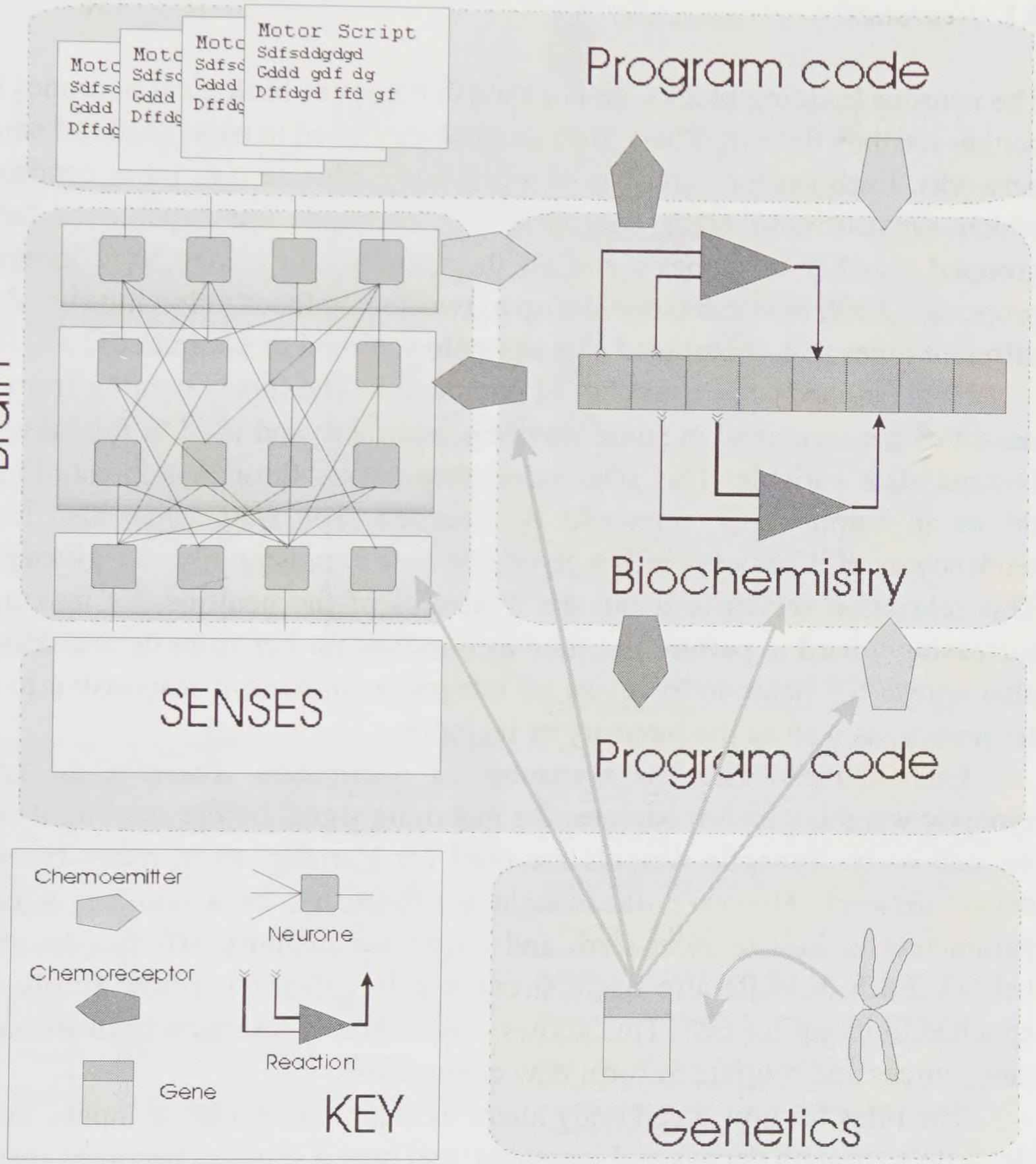


Figure 2. Overall architecture and main components

The environment itself was principally composed of a selection of virtual objects (food, toys, vehicles, etc.), whose properties and behaviour were defined using the meta-language. These objects could be sensed and interacted with by the creature through various generalised messaging systems.

The following sub-sections briefly discuss each of the major components used for constructing creatures, as a preparation for examining how they may be combined to create complex, lifelike behaviour.

4.1 *Neurones*

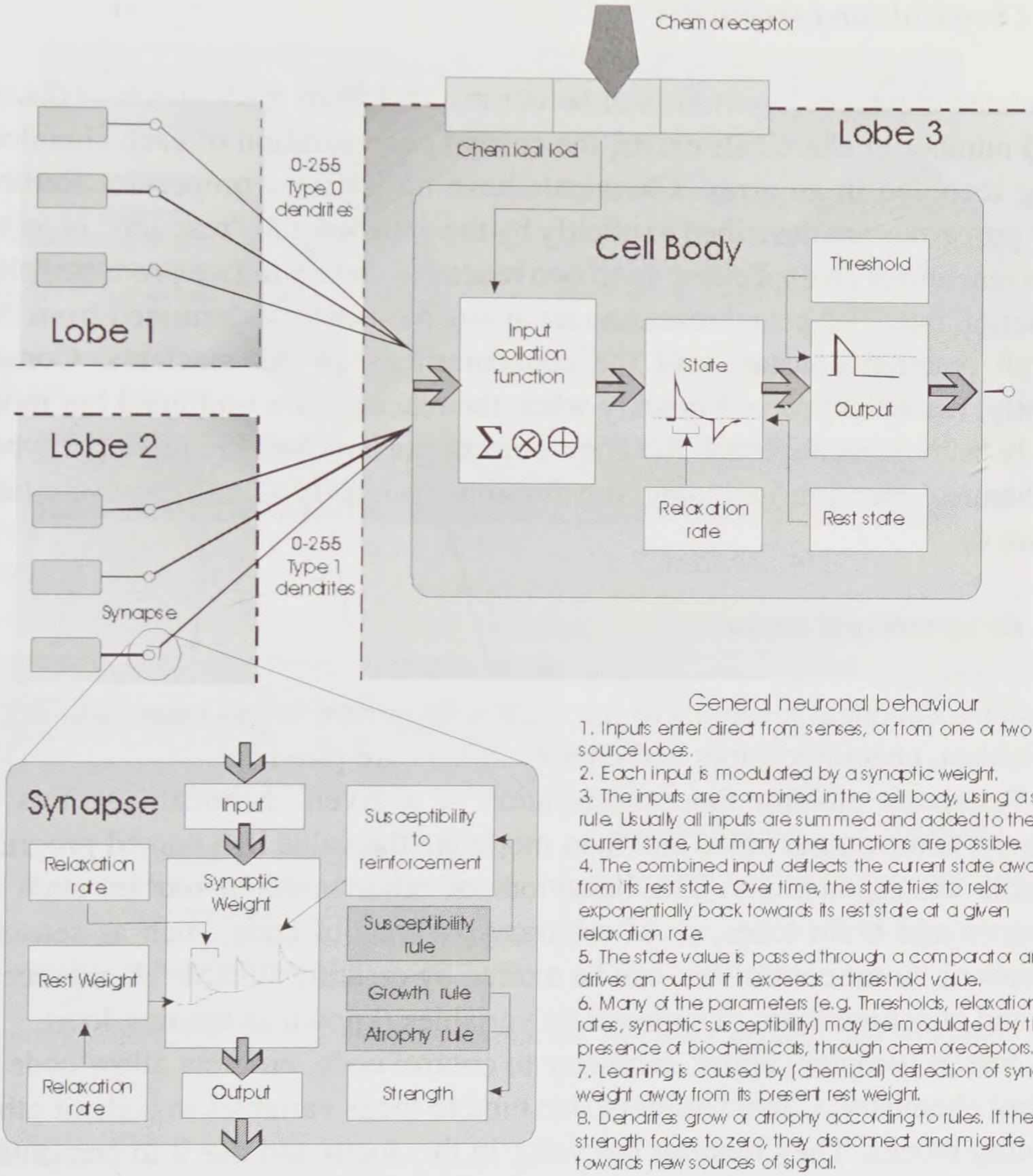
The neurone building blocks are not slavish copies of biological neurones but neither are they the simplified, abstracted objects used in most artificial neural networks. Each neurone consists of a cell body, plus zero or more dendrites, which synapse onto other neurones. The neurones are capable of being grouped together in "lobes", and all the neurones in a lobe have identical properties. Each lobe can innervate up to two source lobes, allowing signals of different types to be combined (for example summed or compared).

Each cell body has a number of parameters. The inputs coming from the dendrites are combined in some way (e.g. summed) and used to modulate an internal state variable. This state value then passes through a threshold and drives an output if the threshold is exceeded. The state value also has a tendency to relax back towards a given rest state exponentially, at a given rate. This relaxation serves to damp the dynamics of the neurone, by making it increasingly hard to perturb the state as it moves further from its rest state. It also allows the neurone to act as an integrator, making it responsive to the frequency, as well as the intensity of its inputs.

Each dendrite also has a number of parameters. There is the usual synaptic weight, which modulates the incoming signal before sending it on to the cell body. Synaptic weights are used for learning, as in many types of neural network. However, the weight is modulated by a number of other parameters to imitate short-term and long-term memory effects (described below). Each dendrite also has a strength value, signifying how firmly it is attached to its source cell. This allows dendrites to disconnect from unimportant sources and migrate to form new connections.

The rules for how a cell body alters its state in response to inputs, how a dendrite's strength decays and increases, and how a synapse becomes susceptible to reinforcement are all programmable, via a simple expression language. Finally, whole lobes of neurones, or individual dendrites, can be made chemically susceptible by attaching chemoreceptors, or chemically active by attaching chemoemitters (see below).

Neurones are thus quite complex and highly configurable objects, from which a wide variety of neural networks can be constructed, each network being made from a heterogeneous collection of lobes (see Figure 3 for a schematic description or Grand, Cliff & Malhotra (1997) for details).



General neuronal behaviour

1. Inputs enter direct from senses, or from one or two source lobes.
2. Each input is modulated by a synaptic weight.
3. The inputs are combined in the cell body, using a state rule. Usually all inputs are summed and added to the current state, but many other functions are possible.
4. The combined input deflects the current state away from its rest state. Over time, the state tries to relax exponentially back towards its rest state at a given relaxation rate.
5. The state value is passed through a comparator and drives an output if it exceeds a threshold value.
6. Many of the parameters (e.g. Thresholds, relaxation rates, synaptic susceptibility) may be modulated by the presence of biochemicals, through chem oreceptors.
7. Learning is caused by (chemical) deflection of synaptic weight away from its present rest weight.
8. Dendrites grow or atrophy according to rules. If their strength fades to zero, they disconnect and migrate towards new sources of signal.

Figure 3. Neurones and their interconnections, showing cell body and synaptic parameters. Note that the detailed interactions between variables are not fixed, but are specified by genetic operations

4.2 *Chemicals and reactions*

Simulated chemical processes can be constructed from reaction networks. A fixed number of chemicals exists, the current concentration of each chemical being recorded in an array. Chemicals have no intrinsic properties. Rather, their properties are described explicitly by the genetics, via "reaction" objects. Each reaction object specifies up to two reactants and up to two products, plus a reaction rate. The actual reaction rate at any moment is determined from the overall reaction constant and the concentrations of the reactants. Consequently, reactions proceed quickly when the reactants are undiluted but more slowly as they get depleted. Reaction objects can simulate the principal types of chemical reaction including substitution, fusion, lysis and catalysis (see Figure 4).

4.3 *Receptors and emitters*

To enable chemical reactions to interface with neurones and other data items (variables), chemoreceptors and chemoemitters are provided.

Receptors monitor the concentration of a given chemical, use this to compute some output value and then modulate the value in a named program variable correspondingly. The behaviour of other building blocks, such as neurones and brain lobes, or of arbitrary portions of code, such as sensory systems or motor controllers, can be altered by creating suitable chemoreceptors and attaching them to appropriate variables (known as binding loci).

Just as receptors allow chemistry to control code, emitters allow code to control chemistry. Chemoemitters also bind to locus variables in code or other building blocks. They monitor the value in this locus and use it to compute a value, which represents the rate at which they secrete a given chemical. For example, a neurone can be made to emit a chemical when it fires, by attaching a chemoemitter to its Output variable.

Both emitters and receptors contain a number of modulation parameters, such as thresholds, scaling factors, nominal values and invertors (see Figure 4). By adjusting these parameters, emitters and receptors can be made to exhibit an assortment of behaviours.

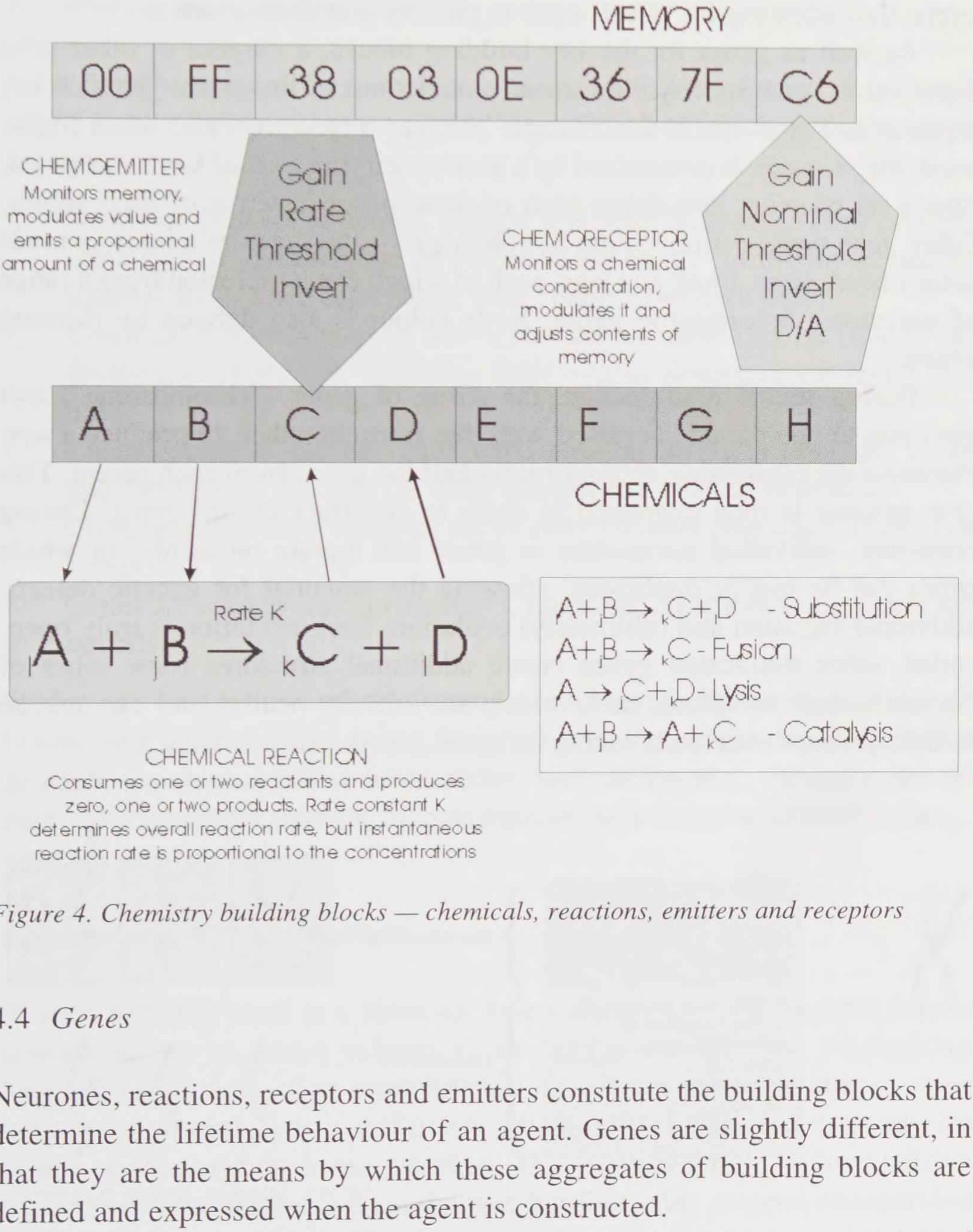


Figure 4. Chemistry building blocks — chemicals, reactions, emitters and receptors

4.4 *Genes*

Neurones, reactions, receptors and emitters constitute the building blocks that determine the lifetime behaviour of an agent. Genes are slightly different, in that they are the means by which these aggregates of building blocks are defined and expressed when the agent is constructed.

Each individual component (reaction, emitter, receptor or lobe of neurones) is defined by a unique gene. Each class of component has its own characteristic gene type, which specifies the parameters for that object. Each gene also has a header, which specifies certain aspects of the gene's expression or modification. Genes can switch on at various stages in a creature's life

cycle, thus allowing for effects such as puberty and senescence.

As well as genes for the key building blocks, a number of other gene types exist. Some specify the various articulations of limbs that generate key postures and motions for the creature. The way a creature walks when frightened, for example, is determined by a gene specifying a set of keyframe poses, plus a set of genes specifying each of these poses in terms of limb angles. Other gene types define simple morphology — the creature's body is made from a head, arms, trunk and legs, each of which can be selected from a range of variations, according to genes. Body colour is also defined by pigment genes.

During sexual reproduction, the string of genes ("chromosome") that gave rise to one parent is crossed with that from the other, to produce a new chromosome comprising approximately half the genes from each parent. This new genome is then expressed in order to construct the offspring. During crossover, individual parameters in genes can mutate randomly, or whole genes can be lost or duplicated, allowing the potential for genetic defects, additional variation and (ultimately) evolution. Such evolution is truly open-ended, since duplicated genes create additional structures (new lobes of neurones, new receptors, etc.), which are initially neutral and can mutate further, perhaps eventually taking on novel functions.

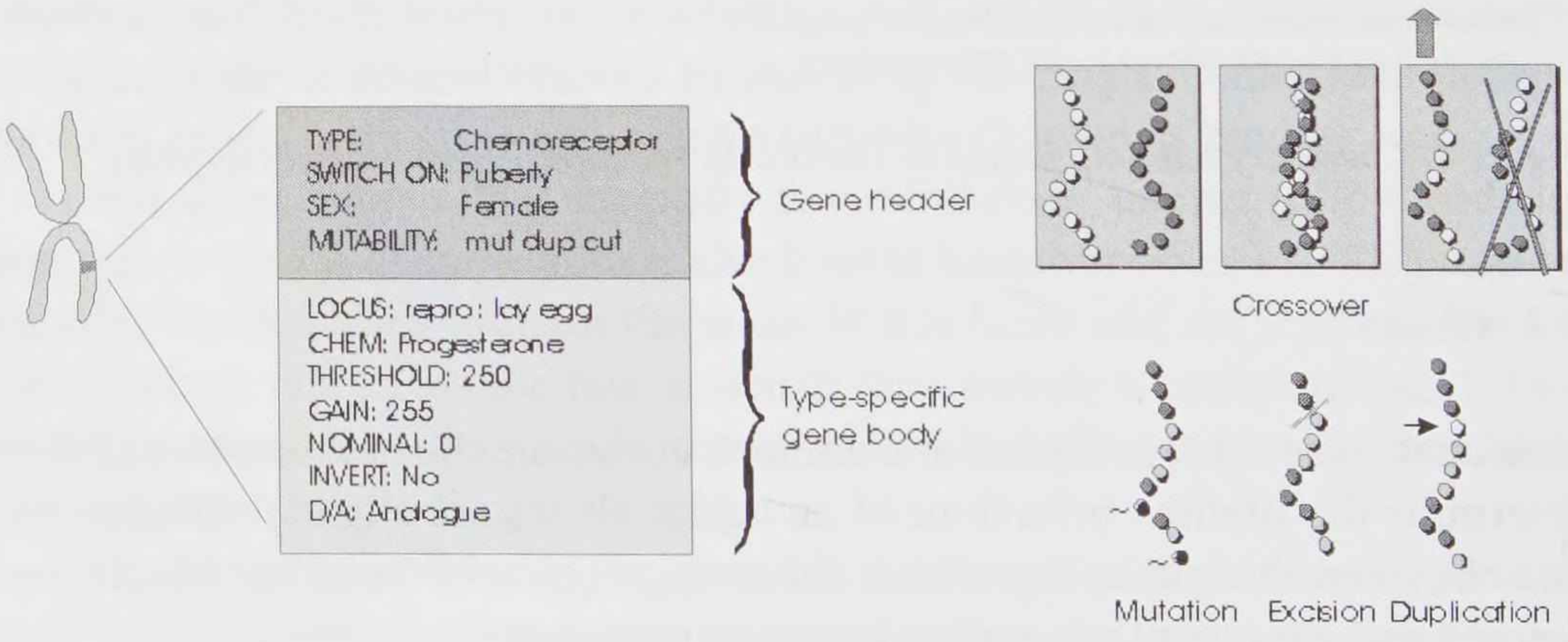


Figure 5. Genetics

5. Putting them all Together

Even though many of these components are quite complex, there are really only a small number of basic building blocks, each of which performs simple functions:

- Neurones connect to senses, actions and other neurones. They combine input signals and compute an output based on the current inputs and the internal state. They can change their connections over time.
- Chemical reactions convert one or two source chemicals into one or two destination chemicals, changing their overall concentration levels over time.
- Chemoreceptors use chemical concentration levels to compute values that can be read by code (or other components such as neurones).
- Chemoemitters read variables set by code (or other components) and emit a chemical in response.
- Genes are used to wire up neurones, reactions, receptors and emitters into circuits.

So, how do we use these relatively simple building blocks to construct socially functioning agents? How do we design chemical and electrical circuits that generate the characteristics listed above, such as learning, curiosity and disease? The following sections discuss each of the key characteristics in turn.

6. Learned Action Selection

In order to justify itself as a 'thinking' being, the creature's brain must connect sensory inputs to motor outputs in an intelligent way. The relationships between sense and action must be built up from a naïve (or almost naïve) initial state at birth, into a competent creature that knows the most sensible course of action to take in any situation. This learning should be susceptible to supervisory reinforcement from the user but must also proceed unsupervised in a relatively complex world, where reward and punishment may lag significantly behind the actions which gave rise to them. During learning, the creature must make rational decisions in the face of novelty and so must be capable of generalising from past experiences.

To see how these aims might be achieved, we can start by visualising a

"perfectly knowledgeable" creature, which knows exactly the best course of action to take in any circumstance. Such a creature could presumably simply have its inputs hard-wired to the appropriate outputs. A 1:1 connection between a sense and an action will not suffice, however, since the most appropriate action depends, not on the individual inputs, but on their combination. For example, the best action for a male creature to take when approached by a female is probably somewhat different from the best action to take when approached by a heavy truck! The sensory fact that "it is approaching" is not sufficient; nor is the single datum "it is female" — it is the unique pairing of the two that defines the relevant sensory situation (schema). So, we can imagine a hypothetical creature in which raw sensory inputs are ANDed together before being sent to the appropriate output, forming sentences such as "IF it_is_a_truck AND it_is_approaching THEN run_away".

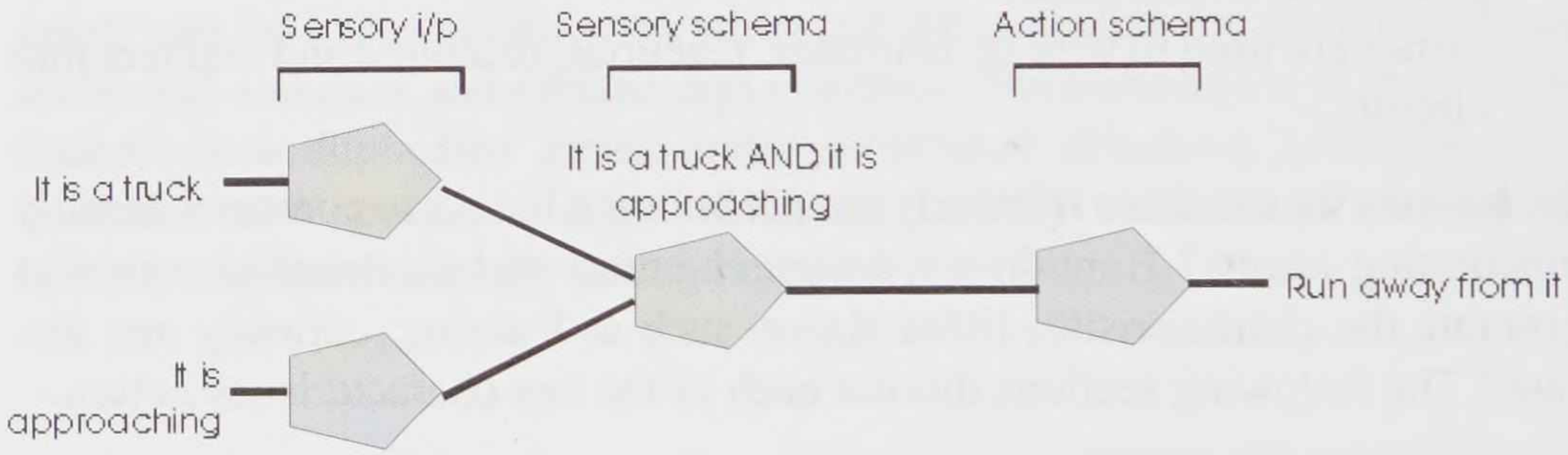


Figure 6. Simple neural rule

So far, so good — the creature is 'intelligent' to the degree that its actions are logically related to its circumstances. However, we do not want or have a perfectly knowledgeable creature — we need one that *learns* these relationships. We need the creature to link sensory combinations to actions in such a way that information about the outcome can be fed back to make such behaviour more or less likely in future. We do not know when such reinforcement is likely to arrive, although we can assume that punishment or reward that occurs shortly after an action has been taken has a greater likelihood of being a consequence of that action than it does if it arrives some time later. So, if we can make our *schema* → *action* connection become *susceptible* to reinforcement, to a degree that decreases over time, then we can simply

reward or punish any susceptible connections at the time such reinforcement comes in, in proportion to the degree of reinforcement multiplied by the current degree of susceptibility. A connection can be considered culpable for causing some painful (or pleasurable) occurrence if it might have been the cause of a recent action. Therefore, a connection should become susceptible if it finds itself conducting a signal (i.e. it relates to a current sensory schema) *and* it connects to an action neurone that is firing (i.e. it may have contributed to the decision to take an action).

Reinforcement is another issue. We must have a mechanism by which a global event (punishment or reward) can make local changes (to synaptic weights) across a wide region (a whole brain lobe). In real brains, such widespread influences appear to be the result of diffuse neurotransmitters or neuromodulators, and in our case we can use the biochemical components to achieve a similar result. If we assume that punishment and reward are represented in the creature by emissions of appropriate chemicals (through mechanisms to be described later), then we can cause our susceptible synapses to be reinforced by attaching two chemoreceptor objects (one for punishment, one for reward) to the appropriate brain lobe. We can then write a rule in the synapses that makes changes to synaptic weight be consequent on the amount of activity in these receptors. At this stage, using chemoreceptors seems unwarranted. However, later we will see how the biochemical building blocks can allow us to generate punishment and reward chemicals in response to drive changes.

So, we now have a mechanism for learning the relationships between sensory schemata and appropriate or inappropriate actions. But what of these sensory schemata? It is immediately obvious that a neural network that can explicitly represent all possible permutations of a moderate number of sensory inputs (a hundred or so) would require a vast number of neurones. Happily, it is most unlikely that any individual creature will experience every one of those possible permutations. Therefore, we need a system in which the creature can form permutations *as they arise*, and can forget them again, should they turn out to have little utility. To allow new schemata to form as new situations arise, we can simply allow our schema neurones to grow towards sources of sensory signal (although we have to make sure that not all neurones synapse onto the first new input combination to arise). Since new situations arise quite frequently, all our neurones will quickly get used up unless we can recycle them. To do this, we can set synaptic strengths to a high value upon

formation, and then define a rule that makes them atrophy over time. When the strength falls to zero, a dendrite becomes disconnected and is free to search for new sources of signal. We do not want all the connections to come loose at the same time but we do need to ensure that there are always sufficient loose neurones to form new connections. So, we need to control the atrophy rate globally across the lobe, to attempt to keep a "pool" of available loose cells. Again, we can model this global behaviour using our chemical building blocks. By attaching chemoemitters to the brain lobe, we can excrete a chemical in proportion to the number of neurones currently with disconnected dendrites. A chemoreceptor tuned to the same chemical can then be used to monitor the concentration of that chemical and hence the number of loose cells. By referring to that receptor in our synaptic atrophy rule we can ensure that synapses only atrophy when the receptor is active, i.e. when the level of the chemical is less than a set threshold. By adjusting this threshold we can control the proportion of loose neurones in the lobe. All that remains is to include a rule that *increases* synaptic strength whenever a sensory schema neurone is used (fires), especially if that firing is followed by some reinforcement, and we have a mechanism for ensuring that unimportant sensory memories get forgotten more quickly than important ones.

Finally, there is one other requirement we must address. When a creature finds itself in a novel situation, it would be most un-lifelike if it either did nothing at all or responded by an apparently random action. It is important that the creature can generalise from previous, related experiences and make rational (even if not correct) responses to novelty. For example, if the creature is rapidly approached by a large truck, and has only previously been approached by large females, it might decide to kiss the truck. This is a very stupid thing to do, but *it is rational* — the creature is generalising from past experience in an explicable, realistic and potentially useful way. To do this, we need to ensure that similar-yet-not-identical circumstances should have some influence on the creature's behaviour, in the absence of any learned response to the specific situation. Happily, for two situations to be regarded as similar, it is reasonable to assume that they must share one or more sensory features. If we make sure (as would naturally happen anyway) that whenever a new sensory situation arises, the network forms the total schema (A+B+C) *plus* all sub-schemata (A, A+B, A+C, B, B+C, C) then the system will automatically work. In our truck example, the situation "it is a truck AND it is approaching" has never been encountered before. However, because the

creature has previously experienced "it is a female AND it is approaching" it will also have formed memories of "it is female" and "it is approaching", both of which will also have been reinforced as a result of the pleasure resulting from the kiss. Therefore there are three positive relationships: "IF approaching AND female THEN kiss", "IF approaching THEN kiss" and "IF female THEN kiss". When the truck scenario arises, two new schemata will form: "IF approaching AND truck THEN ?" and "IF truck THEN ?". The "?" signifies that neither schema yet connects to any action. Both these schemata will be firing, but can recommend no action. However, the schema "IF approaching THEN kiss" will also be firing and so, all other things being equal, the creature will decide to kiss the truck (thus making a rational generalisation about things that approach). When the inevitable happens, the resulting punishment will cause any speculative connection between "IF approaching AND truck THEN kiss" and "IF truck THEN kiss" to be given negative weights, causing such behaviour to be corrected in future. At the same time, "IF approaching THEN kiss" will also be punished. This punishment might counterbalance the positive weight derived from having once kissed a female, and so the solitary fact that something is approaching will be deemed to have no predictive potential. As long as the super-schemata (A+B+C) fire and are reinforced more strongly than the sub-schemata (A, A+B…) the mathematics works out and the creature will both perform and learn correctly.

We have now defined two brain lobes (each defined by a gene, of course). One lobe (large) forms speculative connections to raw sensory data and performs an analogue equivalent of AND on them, to develop a memory of sensory schemata. These memories are strengthened according to their utility and, if unimportant, the neurones will get recycled for re-use via a chemical feedback loop. The other lobe has a small number of cells (one per possible action) but is highly dendritic. These dendrites seek out sources of signal in the sensory schema lobe and are reinforced by punishment or reward through chemoreceptors. The cells have slow relaxation rates, making them integrate the various recommendations and dis-recommendations over a period of time. Whichever neurone is firing most strongly will cause its corresponding action to begin. From this system we have created a 'brain' that is intelligent in its actions, can learn from experiences in a messy asynchronous world, shows short-term and statistical learning, forgets things that were not important and can generalise rational courses of action in novel circumstances.

7. Attention

One more task needs to be built into the creature's brain — we need a mechanism for directing attention. Attention serves several important purposes. Firstly, it minimises the amount of sensory input, by blinkering the creature to all but a single object (plus some general, undirected sensory information). Not only that but it enables the brain to keep all the sensory data about that object bundled together. Without directed attention, the creature would know that some objects were moving and some were making a noise, but might find it difficult to keep track of which object is doing what. Secondly, attention serves to denote the object of a creature's actions. It makes reasonable sense that when a creature chooses to take an action, it should perform the action upon the object it is looking at. Lastly, since attention in these simple creatures equates with gaze, the user can interpret the intentions and interests of the creature by following its gaze. If the gaze can be turned on the user (or the user's 'hand' — the mouse cursor) then the user will feel that the creature is paying attention to her (or not) and will have a valuable source of feedback. If the user can draw the creature's attention to other objects, then yet another line of communication opens up.

Human attention is a complex process, but in our case we can simplify things a great deal. The most important thing is for the creature to pay attention to the object within visible or audible range that is most likely to have meaning at any moment. To a large extent, the object with the most meaning is the one making the most sensorial 'fuss' — the one approaching quickest, the one making the loudest noise, etc. The sensory system can be coded to direct such environmental stimuli into the brain in such a way that different perceived objects are represented by different neurones. In order to control the creature's attention, we can simply set up a lobe of such neurones to act as integrators (by giving them a rest state of zero and a gentle relaxation curve). The neurone representing the object which generates the most extreme or the most frequent stimuli will therefore have the strongest output. Our design therefore allows the creature to be simultaneously aware of all objects in its local environment to the extent that it knows which ones have moved or made a noise, but only be aware of a single object in great detail. Only the detailed knowledge about the current object of attention needs to be fed into the action selection mechanism.

8. Drives and Needs

Our creature is now capable of learning about its environment through reinforcement, but somehow we have to provide that reinforcement. In real life decision-making, there are usually no right answers, only pragmatic ones. In the absence of a measure of perfection, it is hard to quantify the degree of error in a less-than-perfect action. The world is also asynchronous, and the reward or punishment for a course of action may be delayed significantly after the action that caused it. Finally, many decisions will result in a trade-off, in which some aspect of the creature's state improves while some other aspect worsens.

We thus need a mechanism for defining the current 'survival potential' of our creatures, as a measure of the utility of their recent actions. Such a mechanism must take into account the multi-dimensional nature of survival in a 'natural' environment — obtaining food is good, but not at the cost of being wounded. A good way to simulate this multi-dimensional, approximate survival potential is to define a set of needs or drives. For example, we can specify a drive to avoid hunger, which is recorded as a hunger level. The hungrier the creature is, the more pressing its need to reduce hunger. Such a system provides us with our reinforcement signal. Any environmental stimulus that reduces a drive (such as food reducing hunger) must be a Good Thing, and should therefore emit a reward into the brain, positively reinforcing the action that led to that reduction. Anything that increases a drive (for example a wound increases the degree of pain and therefore the drive to reduce pain) should likewise emit a punishment.

Now we can see why it was a good idea to use chemoreceptors in the brain to respond to reinforcement. We can stimulate those receptors by punishment and reward chemicals, and we can derive those chemicals through chemical reactions. The sources for those reactions are other chemicals, which represent drive increasers and drive reducers. As well as generating reinforcers, these reducers and increasers modulate the levels of further chemicals representing the current states of the drives. Last but not least, we can attach chemoreceptors to some of the brain's sensory inputs, which enable the brain to be aware of the levels of these drives, since such knowledge is important for assessing the creature's state and hence the most appropriate actions.

So, an environmental event that decreases the survivability of the creature (say a wound) will trigger the following web of effects. The wound will cause

a pain-increasing chemical to enter the bloodstream. This will react to form two products — more pain chemical (thus increasing the level of pain) and a punishment chemical (which negatively reinforces any recent actions that might have caused the wounding, via the punishment chemoreceptor). Similarly, an event that increases survivability (say a creature enters a warm, sheltered spot) will cause a drive reducer chemical to flow into the bloodstream. This drive reducer must decrease the 'drive to avoid cold' and produce a reward. This can be achieved by defining a reaction in which the drive reducer and some of the drive chemical react together to form reward chemical. Since some of the drive chemical is consumed by the reaction, the drive level will fall.

But why record the absolute levels of the drives at all? Why not simply allow the drive reducer and increaser chemicals to cause reinforcement directly? Firstly, the levels of drives must be monitored by the brain — the creature needs to know how it is feeling while it is choosing an action, since such drives are important motivational information. Secondly, if a drive is reduced to zero, further reductions should not cause reward. Without this constraint, once a creature finds an action that is pleasurable in some way, it will repeat it endlessly. By allowing many stimuli to alter more than one drive, such saturation allows for differential reinforcement. For example, if throwing a ball reduces the creature's boredom drive strongly but also makes it slightly more tired, then the creature will be rewarded for throwing the ball when bored. However, when all boredom has disappeared, it will begin to be punished for the increase in tiredness, thus helping it to distinguish between good and bad times to play ball. Finally, by recording absolute drive levels and allowing them to cause involuntary changes in visible state, such as changes of facial expression, the user is made aware of the mood and needs of her creature.

9. Motivation and Curiosity

Armed with our drives, our attention mechanism and other modules, we only need a few more things to endow our creatures with something approximating to curiosity. Curiosity is an important motivating influence, and our creatures must be motivated to act or they will simply stand there and not take the experimental actions that are necessary for learning to occur.

The boredom drive provides a measure of the need for self-motivation. If

we engineer our boredom chemical to increase in concentration over time, and then decrease sharply whenever the creature performs some activity (especially if that activity results in some entertaining event), then we have a mechanism by which the creatures can learn that playing with objects is a good thing, and standing idle is not.

By making sure that incoming stimuli to the attention lobe are stronger if the object that is causing the stimulation is unfamiliar, we also predispose our creature to take more notice of novel objects. It will abandon a familiar object and seek its boredom reduction from a novel object, should one catch its eye.

The only problem left is how to trigger the creature to act in the first place. After a creature has had a few experiences that show how interacting with the world decreases boredom, it will learn to do more of the same. However, a newborn creature will simply sit there, boredom increasing to saturation, with no incentive to take this first exploratory move.

10. Instinctive Behaviour and Dreaming

To solve this problem we need to introduce a new kind of gene. This gene does not create an object, but modifies a structure — it rewires the brain. What we need to do is give our creatures an 'instinctive' predisposition to 'fidget' when bored. For example, as boredom increases, we want the creature to be less inclined to stand quiescent, so that other courses of action are correspondingly more likely. We cannot hand-wire this behaviour, neurone by neurone, since the brain is too complex (and evolution may have altered its structure). One way to achieve the result is to let the creature 'experience' the situation we want to encourage or discourage, *in utero*. We then want to reward or punish the creature automatically, so that it learns the relationship for itself. An instinct gene, then, disconnects the creature's brain from its environment and feeds it stimuli *as if* the creature was experiencing these things directly. It then forces an action neurone to fire, as if the creature had chosen to take that action in response to those circumstances. Finally it injects some reinforcement chemical and thus causes the sensory schema to link to the action in a positive or negative sense.

Such an instinct gene is biologically rather improbable, although something vaguely similar must exist to pre-program humans to exhibit instinctive (but not simple reflexive) behaviour patterns. However, it is very useful when

used sparingly. Because the instinct works *through* the brain, rather than overruling it, the creature can gradually un-learn instincts through experience. A newborn might, for example, be given an instinct that encourages it to put things in its mouth. This instinct is easily suppressed by negative experience and the creature will quickly learn which things are good for eating and which things are not.

Instinct genes take a long time to process, since learning takes time. Instincts for newborn creatures can be expressed along with most of the other genes during embryology (these creatures hatch from eggs, and incubation provides a good time for processing their instinct genes). However, instincts that do not switch on until later in life (such as a predisposition after puberty to find the opposite sex attractive) would be very useful. Disconnecting the creature's brain to process an instinct gene has to be done during a time when the brain is not needed, obviously, and so a good time to do it is when the creature is 'asleep'. I like to pretend that such somnolent brain activity corresponds to 'dreaming'.

11. A Reason to be Hungry

Many drive changes can be directly related to environmental events or can vary over time (like boredom), but for some this does not seem sufficient. Hunger, for example, should not simply increase steadily over time and be decreased instantly by food; it should rise in relation to the amount of exertion the creature makes, and food should decrease it indirectly, through a digestive process. Happily, the chemical modelling system makes it quite easy to define a digestive and respiratory system for our creatures, and several other advantages then drop out for free.

We can start by making food emit nutrients when eaten and use a network of reactions to convert those nutrients into glucose. To ensure that glucose gets used up at a rate proportional to exertion, we can control its combustion into waste products using a reaction with an enzyme. The concentration of this enzyme can be made proportional to exertion simply by generating it from a chemoemitter, which we can attach to a variable in the code that records the number of joint movements during an animation frame. The more the creature moves, the more enzyme is available and the faster glucose will be burned. Glucose production from food will be slow, but its removal during exertion

may be rapid. To balance these we can introduce a long-term energy store (glycogen) and define a reversible reaction, which stores excess glucose as glycogen or releases it during a glucose debt. This two-way reaction can be utilised in order to control hunger — when glucose is being converted to glycogen, the creature must be digesting, and so we can reduce hunger (by allowing the glucose → glycogen reaction to produce hunger-reducer as a by-product). Similarly, we can increase hunger when glycogen is being converted to glucose, since this signals a blood sugar debt. Lastly, we can assume that a creature that has completely exhausted its long term energy store is no longer viable, and so by attaching a glycogen chemoreceptor to a flag in the code, we can dictate that creatures 'die' when their glycogen levels fall to zero.

12. Disease

Note that this biologically plausible mechanism for controlling hunger also allows us to interfere with the process in realistic ways. We can now define toxins that can be ingested in the same way as nutrients, but which interfere with the digestive process and endanger the life of the creature. By placing such toxins in the environment (for example in herbs), we can ensure that the user has an incentive to protect her creatures.

Another source of disease is also possible. If we define some 'bacterium' objects in the environment, which can infect a creature and emit toxins into its bloodstream, then the creature can catch cold. By defining a 'histamine' chemical and adding histamine receptors that invoke sneezing and coughing, we can ensure that bacteria can be passed on contagiously from creature to creature. Since bacteria that produce histamine will be preferentially transmitted, there is a survival advantage for those types of bacteria. All we have to do is make the chemical emissions from different bacteria 'genetic' and mutable, and we have the possibility of evolving populations of pathogens!

To protect our creatures from these pathogens (and incidentally create a co-evolutionary scenario), we could demand that all bacteria, as well as emitting assorted chemicals that might act as toxins, also contain chemicals that behave like antigens. By defining an equivalent set of antibody chemicals in our creatures and setting up reactions that create appropriate antibodies in the presence of antigens, we can define a system that allows creatures to build up (or inherit) immunity to different classes of bacteria.

Creatures will now have an in-built tendency to heal automatically, if infected. However, some toxins may act too quickly, and we can also require that antibody production is reduced in weakened creatures. Having thus toyed with the emotions of our users by making creatures that can fall ill, sicken and die, we can redress the balance in an absorbing way by littering various 'medicinal herb' objects around the environment. These herbs can be designed to emit chemicals that negate the effects of toxins or reduce their symptoms (antihistamine, say). By learning which herbs do what, our users can learn how better to take care of their creatures. Our experience shows that dealing with sick creatures has immense power to generate emotional involvement from the user, which is, after all, the object of the exercise.

13. Sex and Reproduction

Another natural and expected lifelike phenomenon is sexual behaviour, leading to reproduction. In the Creatures game we had to be fairly circumspect about this, but the notion of 'family' is a vital aspect of social creatures and so it was decided to incorporate it. Since the complete description from which a creature is assembled resides in its genes, the general mechanism for reproduction comes almost for free. By taking the genome from each parent and crossing them over (see Figure 5), we can produce an offspring that inherits some genetic features from its mother and some from its father. Because the relationship between genotype and phenotype is not trivial, the ways in which offspring resemble their parents can be very subtle (and provokes a great deal of interest — even pride — from their owners). Some offspring are likely to suffer from inherited or mutant diseases, of course, but in terms of promoting social interactions between agent and user, this is no bad thing.

Reproduction is not all that sex is about. From a sociological aspect, sexual behaviour is even more important. A good start towards invoking sexual behaviour is to mimic the various hormonal influences that occur in males and females. A sex drive chemical can be implemented in the same way as the hunger drive, but we can define a crucial part of this mechanism so that it will not operate until a certain gene gets expressed at the age of puberty. Other pubertal genes can initiate hormone cycles. In males, we can define testosterone, have it modulated by environmental cues (for example in response to a female 'pheromone') and let it affect the sex drive. Male hormones

are relatively simple. For females, it would be good to have some kind of ovulatory cycle and, happily, this can be created by attaching an oestrogen chemoemitter back-to-back with a chemoreceptor, to create an oscillator. The 'Oestral cycle' that results can be used (via another chemoreceptor) to control ovulation. We can set it up so that females are fertile only during certain parts of their cycle. When a female conceives (receives a genome from a male), we can control the progress of pregnancy with hormones, too. An 'I am pregnant' locus can hold a chemoemitter, which generates a rapid increase in a hormone that suppresses ovulation (by consuming oestrogen). Another emitter can generate a slower release of a hormone, which builds up steadily until it reaches the threshold of a chemoreceptor connected to the 'lay an egg' locus.

When the child is born, any parental interest, sibling rivalry and suchlike are up to the creatures themselves to discover, although we can help things along by firstly making sure that creatures can recognise their immediate relatives (by having 'it is my parent', 'it is my sibling', etc. sensory inputs). Secondly, we can use instinct genes to encourage children to keep close to their parents and, of course, to encourage members of the opposite sex to find each other repulsive or attractive, as their age requires.

14. Other Social Interactions

When a creature discovers another creature, they have a natural tendency to huddle together and interact with each other excessively. This is because creatures make a deal more 'fuss' than most environmental objects and so grab other creatures' attention. Various interesting feedback loops can also arise. If two creatures discover that they can excite mutual pleasure from stroking each other on the head (the method by which we allow the user to reward a creature for good behaviour) then they can quickly become obsessed by the activity. To compensate somewhat for these features, we need to introduce a new drive. By establishing the concept of 'overcrowdedness' and making the nearby presence of other creatures increase this drive, we can ensure that too much socialising is seen as a Bad Thing. An instinctive response to walk away from creatures when overcrowded can help to ensure the right lessons are learned. If we have overcrowdedness, then we should also have loneliness, and the balance of these two drives can help control the degree to which creatures band together. Similarly, a healthy amount of anger drive can compensate for

the pleasure of 'stroking' interactions with other creatures by making over-repetition of it lead to irritation.

15. Language

Social interaction between the creature and its owner can be difficult, since virtual creatures are not tactile and the user does not feel part of the creature's world. In compensation, a simple language mechanism was introduced, part neural, part code. Creatures can learn verbs, representing the action schemata they are capable of expressing, and nouns, representing the names of objects in their world. To make the game international and natural, it was decided that the creatures should have to learn these words, with help from the user. In the case of nouns, it is quite easy — the user simply has to attract the creature's attention to an object (for example by waving it around) and speak (type) its name. The creature will then learn to associate the word with the class of object. Verbs are trickier, but this problem was solved by placing a small 'educational computer' in the world, in which verbs are represented as short animations (a beckoning stick figure represents the 'approach it' schema, for example). By typing words into the computer, the user can teach her creatures the word for each action.

When a creature hears words that it has learned, these pass as sensory data into the neural network, where they can invoke a response. Commonly, the user will type in a verb-noun sentence, such as "press the button". The noun will be sent to the attention lobe and will encourage the creature to pay attention to any buttons it is aware of. The verb will be sent to the action selection system, where an instinct gene has previously wired up a positive relationship between each verb and its corresponding action. As a consequence, the creature is likely to do as it was told, although poorly learned words, other pressing commitments or a learned aversion to following the user's orders can easily override the command.

Finally, words can be used by the creature to explain to the user what it is doing ("Terry get carrot") or how it feels ("Terry hot"), again by using the available repertoire of taught verbs and nouns ("Terry" is, of course, the noun that our hypothetical user associated with the creature itself).

16. Conclusions

In this paper I have tried to show how quite subtle and complex intelligent and social behaviour can be created in a synthetic agent constructed largely from circuits of relatively simple building blocks, wired together in biologically plausible ways. Such notions as 'curiosity', 'illness' and 'generalisation' are neither explicitly programmed into the creature as a whole nor into any individual part of it. Hence, I think it is reasonable to consider these phenomena to be emergent properties. In Complexity Theory circles, emergence is sometimes regarded as something rather ineffable, as if to be emergent a phenomenon must *in principle* not be predictable from the properties of the parts alone. I prefer to think of emergent phenomena as properties of wholes that are not *immanent* in the parts, but which can still be understood and indeed deliberately engineered into existence. I hope that the work described above has shown that it is possible to engineer quite subtle phenomena, given suitable building blocks.

The concept of wiring together building blocks to build circuits is subtly but (I believe) crucially different from that of stringing together computer instructions to form algorithms. The latter is serial and procedural, while the former is parallel and structural. Many lifelike phenomena are most convincing as emergent properties, and high-level, creative intelligence is necessarily emergent (you cannot describe behaviour as creative if it was explicitly preprogrammed). Emergence is, almost by definition, characteristic of systems with multiple interacting parts, and I think that a parallel, connectionist approach is therefore important in AI. Obviously, digital computers are inherently serial, procedural devices and everything that uses them must ultimately be implemented as an algorithm. Nevertheless, simulated neurones, biochemicals and suchlike can be considered as *virtual* parallel machines, even if they are simulated on a serial computer. Such second-order structures (Grand 1997) can be achieved by simply time-slicing the program, so that events which occur within the same time slice are considered to be simultaneous. Intelligence and lifelike behaviour may only be possible in parallel systems, even if that parallelism is a sham. Such pseudo-parallelism is clearly implemented as a serial algorithm, and hence algorithms are not prohibited from exhibiting intelligence. However, it may be that intelligent behaviour is a property only of the class of algorithms that contain a time-slicing loop. If so, to pursue the search for intelligent, lifelike, perhaps ultimately conscious

artificial agents by any other means may be futile. At the very least it may be inefficient, just as a spreadsheet (a pseudo-parallel system) is more efficient and 'natural' to use than a (serial, procedural) programming language for certain kinds of computational task, even though all spreadsheet applications are themselves written in such a language.

Notes

1. Published by Mindscape Inc. in North America and Warner Interactive in Europe. Copyright © Cyberlife Technology Ltd. 1997.

References

Blumberg, B. 1994. Action Selection in Hamsterdam: Lessons from Ethology. In Cliff et al. 1994.

Cliff, D., P. Husbands, J.-A. Meyer & S. W. Wilson (Eds.). 1994. *From Animals to Animats 3: Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior (SAB94).* Cambridge, MA: MIT Press.

Grand, S. 1997. Three observations that changed my life. *IEEE Expert, Intelligent Systems and their applications 12(6).* Los Alamitos, CA: IEEE Computer Society.

Grand, S., D. Cliff & A. Malhotra. 1997. Creatures: Artificial Life autonomous software agents for home entertainment. In W. L. Johnson & B. Hayes-Roth (Eds.), *Proceedings of the First International Conference on Autonomous Agents.* New York: ACM Press.

Hayes-Roth, B. & R. van Gent. 1997. Story-Making with Improvisational Puppets. In Johnson & Hayes-Roth (Eds.) 1997.

Loyall, A. B. and J. Bates. 1997. Personality-Rich Believable Agents That Use Language. In Johnson and Hayes-Roth (Eds.) 1997.

Sims, K. 1994. Evolving 3D Morphology and Behavior by Competition. In R. Brooks & P. Maes (Eds.), *AlifeIV: Proceedings of the Artificial Life IV Workshop.* Cambridge, MA: MIT Press.

CHAPTER FOURTEEN

Machine-Mediated Communication

Agents of Representation

Bill Vorn

1. Robotic Art

I have been involved with robotics since 1992 working with L.P. Demers on many different interactive robotic installations on a creative and artistic level.[1] We use robotics and multimedia to pursue our research on reactive environments and the embodiment of life in inert materials. Our goal is to present robotic machines not as virtuoso specialized automatons but rather as expressive animated works of art. We also explore the reformulation of sound and light applications by simulating organic and metabolic functions and by creating dynamic virtual architectures.

Since our very first project, *Espace Vectoriel* (eight robotic tubes moving freely, reacting to the viewers and projecting sound and light in a directional way - see Figure 1), and without any conscious intent, our creative work has evolved from abstract shapes (tubes, hemispheres, cubes) to more concrete structures (bodies with arms or legs). Even if our goal is not to imitate living creatures like animals or humans, our machines are initially conceived with metaphorical behaviors based on real life ethology.

Our latest piece, *La Cour des Miracles*, is both an attempt to express the profound human nature of the machinic realm as well as the profound machinic nature of humankind. By assuming the role of miserable characters, the machines are no longer simple animated objects but elevated to the rank of actors, i.e. agents of expression. In a world populated by these actors, social

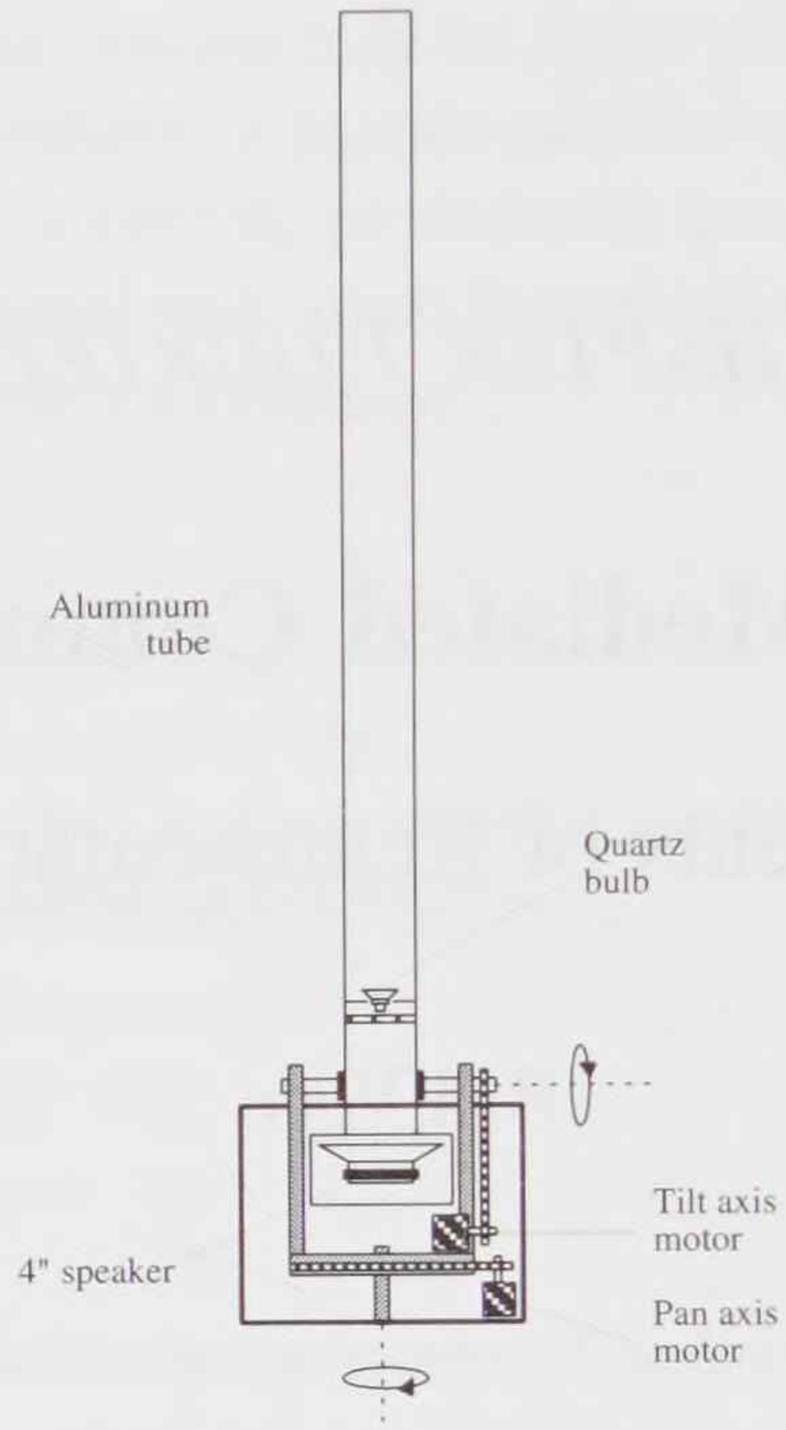


Figure 1. Espace Vectoriel's robotic tube (© L.P. Demers & B. Vorn 1992)

interactions of a different complexity become possible (between robots, between robots and viewers, and between the viewers themselves).

We consider the evolution of our work toward more biomorphic machines to be the result of an always greater need to impress viewers, especially on the sensory level. This evolution emerges both from our own will to communicate something and from the feedback we receive from the public. These artworks seem to generate more symbolic interpretations when their rendering reaches a less abstract condition.

2. Artificial Life as Media

In parallel to my interest in robotic art, part of my academic life is devoted to the study of Artificial Life (AL) as a communication phenomenon. My current research project seeks to reveal the emergence of automatic processes in real-life and screen-based computer-mediated communications that are becoming a new vector for carrying information and meaning between people. This

project will look at machine-mediated messages from both user's and creator's points of view, and through the individual and social behaviors engendered by the inner interactions. It is both a quantitative (statistical / topological) and qualitative (typological) approach to the study of message and communication automation.

I suggest that robotics (real and virtual) should be considered not only an industrial, scientific or even artistic medium (Kac 1997), but also as a media of communication. Even if they generally interact with the audience on a one-to-one basis, robotic agents can carry messages and informational content as efficiently as traditional mass media. Message automation turns hardware and software robots into communication agents, and the instantaneousness of digital information may soon rise to the critical level, creating mass mediation. For McLuhan (1968), automation brings true mass production, not in terms of quantity, but in terms of global instant binding.

3. Artificial Life Survey

This section describes the methodology used in the current stage of the Artificial Life as Media research project.

One of the best ways to find out the effect of a media on its viewers (or users) is simply to ask them what they think, I therefore chose questionnaires as a tool for data acquisition. The use of questionnaires also allows me to collect a certain amount of quantifiable data on which to perform statistical analysis. The Artificial Life Survey is a series of questionnaires geared to the general public and to expert users who have some experience interacting with AL cybernetic organisms.

The project uses six different questionnaires, corresponding to six actualization categories related to Artificial Life: agents, bots, computer viruses, AL simulations, Tamagotchis and robots. The choice of these categories is quite arbitrary (there is no universal classification) but they do cover the most common artifacts produced by AL. As this study is principally about interactive automatons, some categories (like cellular automatons and genetic algorithms) have been deliberately excluded. There is no simple and commonly-accepted definition of each category (for example, what is an agent?), so each respondent had to answer according to his or her own definition of that particular category.

While these questionnaires were intended both for the general public and for expert users, this approach has certain biases. Firstly, experts and specialists were far easier to reach as they tend to belong to well-identified electronic communities. Additionally, it is possible that some people from the general public might have failed to respond because they thought that these questionnaires were intended for experts and experienced users only.

Each questionnaire was composed of about forty questions. The questions had multiple choice or discrete value answers (on a scale of 1 to 10). Some questions were common to all the categories; others treated a specific category. The questionnaire was anonymous but there were some demographic questions (sex, age groups, level of education, etc.) to allow further classification and analysis. Rather than being printed and distributed by mail, the questionnaires were available on the Internet. Each questionnaire was an HTML page where users could reply using a radio button interface (a single choice was then required for each question). For geopolitical reasons, all questionnaires were written in both French and English versions.

The URLs of the questionnaires were published in most Internet search engines, related newsgroups, email discussion lists and electronic newsletters. Obviously, the survey was restricted to people with Internet access and to users with some interest in one of the predefined categories, but this constraint made more specific questions possible for each questionnaire. From November 26th, 1997, until March 1st, 1998, a total of 482 responses were received for all six categories. 38 questionnaires were discarded because they were left blank or incomplete. 10 questionnaires were discarded because of multiple replies and redundancy. 434 questionnaires were kept for analysis. Table 1 shows the distribution of respondents by category. More detailed results are available on the Internet.[2]

Table 1. Distribution of respondents by category

Category	*Respondents*	
Tamagotchis	155	36%
Alife Simulations	78	18%
Robots	63	15%
Bots	59	14%
Viruses	48	11%
Agents	31	7%
	434	

4. Categories and Results

In order to find similarities or differences between artificial lifeforms, six categories were intuitively defined: agents, bots, computer viruses, AL simulations, robots and Tamagotchis. Some categories may overlap onto others; the categories should be considered as guidelines rather than boundaries. The results, still at an early stage of examination, are presented here for each category in narrative form and in tables.

4.1 *Agents*

Even if the term "agent" may be used to describe any AL creature (real or virtual) from robots to viruses, the questionnaire uses this term to refer to what is generally known as "autonomous intelligent mobile agents." An agent is defined as software that assists people, acts on their behalf and solves problems like information overload (Gilbert 1995). An extensive overview of the most common definitions of this term can be found in Franklin and Graesser's taxonomy (1996). Unfortunately, as agents are highly specialized and often experimental entities (many intelligent agents, like Telescript agents, exist only in relatively closed systems), very few people interact directly with them or even know they exist. Only 7% of all respondents were interested in this category.

Of these respondents, the majority have interacted with information gathering and processing agents (65%), mostly on the World Wide Web (58%). "Behavior" is the characteristic most frequently cited by respondents (42%) as typifying the agent they used or interacted with. "Autonomy" and

Table 2. Percentage of preference for agent and robot properties

Property	*Agents*	*Robots*
autonomous	26%	27%
learning	29%	17%
communicative	10%	10%
mobile	10%	8%
reactive	6%	14%
flexible	6%	14%
goal-oriented	10%	3%
character	3%	6%
temp. continuous	0%	0%

"learning" were the most preferred agents' properties, a preference expressed by respondents in the robots category as well. Table 2 shows a comparison between these two categories and users' preferences for the nine properties proposed by Franklin and Graesser to subclassify agents.

Speech recognition and understanding is the most important feature requiring improvement in the creation of better agents, as 52% of respondents chose that characteristic. Table 3 shows preference for the same feature in other categories.

Finally, 61% of respondents replied that agents have somehow changed the way they see the world and 74% believe that agents do carry (to some extent) an informational content (messages, ideas, cultural values, etc.).

4.2 *Bots*

This category consists of most types of software robots (cancelbots, chatterbots, crashbots, eggdropbots, gamebots, floodbots, knowbots, mailbots, spambots, etc.) that do not fall into the autonomous intelligent mobile agents category. Bots are computer programs (like CGI scripts) which act as automatic functions or software automatons. They are not as mobile and autonomous as agents claim to be, but they do act on behalf of their creator and often bear some anthropomorphic characteristics. While some scientists and programmers concentrate their efforts on the technologies and rhetoric of agents in search of new ways to achieve problem solving and machine intelligence, others flood cyberspace with these virtual robots in a more intuitive way, concerned only with the bots' actions and their effects on the users.

Although Leonard (1997) makes a distinction between agents and bots, claiming that agents do not require the accoutrements of personality (like

Table 3. Distribution of "speech recognition" as the most important feature to improve

Category	*Respondents*
Agents	52%
Bots	47%
Robots	30%
Tamagotchis	26%
Alife Simulations	----
Viruses	----

human names, the ability to crack a bad joke, etc.), the current survey reveals that the percentage of respondents who consider the visual aspect of agents to be a formal representation (a person, an animal or a realistic drawing) is higher than it is for bots. Table 4 shows the importance of visual appearance for each category.

As we can see, computer viruses, agents and bots do not really express themselves through visual forms of appearance (abstract or representative). Compared to Tamagotchis and AL simulations which manifest themselves to a large extent through representation, they are still mostly perceived to be virtual and abstract as computer code.

Even though there is a wide variety of bots, knowbots (31%) and chatterbots (29%) are the most popular forms, and the World Wide Web is where they can usually be found (46%). Most people (59%) think that bots are the result of researches in Artificial Intelligence (as opposed to Robotics, AL or Cybernetics), and the majority (68%) think that bots are tools, rather than games, experiments or simple automatons.

Only 27% of respondents replied that bots have somehow changed the way they see the world but 54% believe that they do carry an informational content.

4.3 *Computer viruses*

Computer viruses (viruses, trojan horses, worms, etc.) are in a different category because they are probably the most lifelike artificial lifeforms as well as the most harmful. While a computer virus does bear all nine properties of a living organism stated by Farmer and Belin (1992),[3] it may be shortly defined as "a piece of code with two characteristics: (1) At least a partially automated capability to reproduce. (2) A method of transfer which is dependent on its

Table 4. Distribution of visual appearance by category

Category	*Representation*	*Abstraction*	*Nothing at all*
Alife Simulations	65%	12%	23%
Tamagotchis	62%	26%	12%
Robots	40%	22%	38%
Agents	29%	13%	58%
Bots	17%	10%	73%
Viruses	10%	4%	85%

ability to attach itself to other computer entities (programs, disk sectors, data files, etc.) that move between these systems." (Stubbs and Hoffman 1990)

According to 44% of respondents, the most common type of computer virus is the boot infector virus and diskettes are the most common infection vector (73%). 75% of respondents think that the virus metaphor is quite accurate to describe this kind of software. While 63% of people believe viruses to be rather harmful or very harmful, nearly as many (60%) think that they make computer systems and networks evolve. Of all the properties Farmer and Belin use to define a living organism, self-reproduction (46%) seems to be the most appropriate one to describe computer viruses.

48% of respondents replied that computer viruses have somehow changed their perception of the world and 54% believe that they do carry (to some extent) an informational content.

4.4 *Artificial Life simulations*

This category includes closed systems and environments created to experiment with AL principles and to evolve artificial creatures and phenomena (SimCity, SimLife, SimEarth, Tierra, Technosphere, Creatures, etc.). Agents in AL simulations are software objects that interact with each other and respond to parameter changes induced by the user. Even if these simulations often look like games or pastimes, 78% of respondents claimed that they are mostly tools, social or scientific experiments. 68% of respondents had the impression that the simulation was somehow behaving like a living system, but 73% found that it was unrealistic.

54% of respondents stated that AL simulations have somehow changed their vision of the world. A strong percentage of respondents (74%) found that simulations carry an informational content such as ideas and cultural values. This can perhaps be explained by the means AL simulations use to represent reality: biomorphic visual elements (plants, animals, humans) and "natural" agent behavior (birth, evolution, death). Even if most of these programs show some profound Darwinism in their basic evolutionist concept, only 4% of respondents said that "Survival belongs to the strongest" was the message that describes them best. Among those proposed, the hidden message people preferred most (35% of respondents) was "The whole is greater than the sum of the parts."

4.5 *Robots*

Robots are the tangible manifestation of AL in the real world as they bear most of its properties (except self-reproduction). For Levy (1992), real Artificial Life is robotics. Robots are not only a virtual model (a pattern in space and time) but also a dynamic and evolving phenomenon embodied in matter (Demers and Vorn 1995). All mechanical and physical robots are included in this category without any particular distinction (industrial and educational robots, theme park robots, competition robots, robotic installations and sculptures, etc.).

Even if robots are true descendants of the mechanical automatons of Vaucanson, Maillardet and the Jacquet-Droz family (Cohen 1968), they now have a different role to play than they had in the 18th century. Not surprisingly, 78% of respondents actually think that robots are mostly tools, rather than experiments, games or artworks.

According to 40% of the respondents, vision is the most important sensorial feature of a robot. This makes sense if we consider that 62% of respondents think that robots are living in the same world as we are and not in their own abstract world. Autonomy (27%) and learning (17%) are the most preferred properties of a robot.

Strangely, most people (38%) think that robots do not look like anything in particular (a human, an animal, or even an abstract shape).

63% of respondents replied that robots have somehow changed something in the way their perception of the world and 63% think that they do carry (to some extent) an informational content.

4.6 *Tamagotchis*

This category is not commensurate with the others because Tamagotchis are commercial products and sold as children's toys. But Tamagotchis do hold our interest and can be considered as the first mass production of an AL creature and the first "mass-mediatization" of AL in the general public (40 million units of the original version were sold worldwide in 1997). In contrast to all other categories, the majority of Tamagotchi users are young people (56% under 25) and mostly females (58%, see Table 5).

Table 5. Distribution of female respondents and respondents under 25

Category	*Under 25*	*Females*
Tamagotchis	56%	58%
Alife Simulations	44%	24%
Viruses	40%	19%
Bots	36%	22%
Agents	32%	16%
Robots	22%	6%

48% of respondents think that Tamagotchis look like animals rather than like realistic or abstract drawings and 55% think that Tamagotchis are living in the same world with others of their kind. Maybe this is why 37% of respondents would like Tamagotchis to be able to communicate with each other, as a preferred feature to be added for a better toy. Surprisingly, when asked about what they learned from owning a Tamagotchi, only 1% of respondents selected the Darwinian answer ("Survival belongs to the strongest"), compared to 26% who replied they did not learn anything at all.

Only 20% of respondents replied that Tamagotchis have somehow changed their perception of the world but 52% are thinking that they do carry (to some extent) an informational content.

4.7 *Preliminary overall results*

One of the most surprising results of this survey was a contradiction: despite a widespread perception that AL manifestations do carry an informational content (messages, ideas, cultural values), most people do not think that these productions qualify as media. In fact, 60% of all respondents think that some kind of content is conveyed even though 68% do not think that AL creatures are mediating vehicles of meaning (see Table 6).

This contradiction may be due to the general perception of the mass media themselves. Television, film, radio and even the Internet are very different from automatons in terms of autonomy, temporality and relationship with the audience. If some softbots can behave like directional mass media (spambots, for example), most of them interact with users on an individual basis. However, individuality is still another impression caused by projection and anthropomorphism.

Table 6. Percentage of positive answers to "does carry content" vs. percentage of negative answers to "is a media"

Category	Respondents A		Respondents B	
Alife Simulations	58/78	74%	50/78	64%
Agents	23/31	74%	19/31	61%
Robots	40/63	63%	36/63	57%
Viruses	26/48	54%	38/48	79%
Bots	32/59	54%	36/59	61%
Tamagotchis	80/155	52%	116/155	75%
	259/434	60%	295/434	68%

Finally, after preliminary analysis, the survey reveals another interesting point regarding users' preferences. In a general manner, the term "behavior" was the most frequent answer to the "What struck you most in the current category?" question. Compared to other possible answers (function, look, personality, autonomy, nothing special, etc.), more than 37% of all respondents chose that specific feature (see Table 7).

5. Behavior as an Ontological Framework

In the fields of traditional philosophy and metaphysics, the term "ontology" usually refers to the nature of being and the subject of existence. In Artificial Intelligence, knowledge sharing and software agent technology, an ontology is often defined as the "specification of a conceptualization" (Gruber 1993). An ontology describes how the world in which the agent lives is constructed,

Table 7. Behavior as percentage of answers to the "What struck you most?" question

Category	*Respondents*	
Viruses	26/48	54%
Alife Simulations	39/78	50%
Agents	13/31	42%
Robots	25/63	50%
Tamagotchis	43/155	28%
Bots	13/59	22%
	159/434	37%

how this world is perceived by the agent and how the agent may act upon its world. Robotic ontology is generally defined by a formal description of the component parts and their individual role and behavior.

Behavior is a key word in automaton design and actualization. A certain level of realism may be achieved by the illusions induced by actions and reactions of the animats and the machines: the success of this dynamic form of computer-mediated communication may be measured by the effectiveness of the simulacrum. An effective simulation of the living is the result of different parameters acting as impression triggers (visual appearance, sound emission or physical movement, for example), but behavior may be seen as the most convincing one as it gives a strong impression of autonomy and self-consciousness.

Uncertainty also plays an important role in the behavioral relation with the viewer. Animated metal parts in a robot or dots on a computer screen can be seen as being alive if they move and react in a non-repetitive and unforseeable way, giving a strong impression of self-decision and autonomy. AL creatures do not have to be necessarily figurative representations (anthropomorphic or zoomorphic) to be convincing, as long as they manifest autonomous behavior in the interaction process.

6. Simulacra and Simulations

The strength of the AL simulacrum is emphasized by stimulating an unavoidable reflex of anthropomorphism and projection of one's internal sensations onto the outside world. Suspending deliberately his disbelief, the viewer generally accepts this illusion as a natural thing, convinced mostly by changing movements in his visual and acoustic fields of perception and reactions engendered by his own actions. Transforming the viewer into a stimulus also transforms his immediate relationship with the animats and machines (Demers and Vorn 1998).

A good example of this phenomenon is our robotic installation *La Cour des Miracles* where robots are nothing but more or less abstract metal shapes. These machines have no skin, no head, no eyes and no ears. They have some articulated arms that do not bear any great resemblance to organic ones. Even though, here is a comment one viewer made after visiting the installation:

> [...] Although I'm not really a robot lover, there was something about these metallic skeletal pieces, caught in cages, chained to walls, freakishly dismembered, screaming and writhing their agony that engaged me despite my prejudice. The work played at an edge of human-machine that, thanks especially to the sounds —whispering, howling, groan-like — and to the pained distortions of the movements, evoked a disturbing border state that much of the cyborg-mania misses. These miraculous/ horrific, simple and strange machine freaks expressed and evoked an alienation from the smooth high tech control-desire of the computer world as well as suggesting the impossibility of escape.[4]

In this installation work, all behaviors (movements and expressions) were initially implanted with the idea of creating specific characters: the machines were built in such a way that they were perceived as being dysfunctional organisms, almost like monsters and freaks. They were intentionally designed to look like unusual and untamed creatures as they showed erratic and abnormal behaviors. For example, the "Convulsive Machine" (see Figure 2) emitted piercing sounds and shook its thin metallic structure with frequent but irregular spasms when viewers walked by. Movements were amplified by light beams and convulsions by metal springs. The resulting effect was a disturbing impression of helpless animality.

For Penny (1997), agents work only because they trigger associations in the user. This is true and undeniable as long as we do not reduce an agent to the single state of sign or symbol. Agents are obviously signs and symbols, but they are dynamic messages. They are signs that have the possibility to change their own value as they act, evolve and react to their environment. AL creatures and agents create signification when interacting with the user or among themselves. This is why social agents and artificial social systems manifest different meanings when we examine them from a local or global point of view.

Until now, the existence of AL agents has been the result and expression of their internal program, a script or set of rules initially written by a human hand. Even in the most abstract cases, these programs carry the cultural memes of their authors; agents become the expression of their personal intentions. In a closed system like a computer simulation true and fully spontaneous emergence of an autonomous behavior is quite improbable as the basic rules are always predefined in an arbitrary way then refined according to one's goal. In fact, there is always an idea of finality in the conceptual premises of an automaton. Even in the case of artificial social systems, social

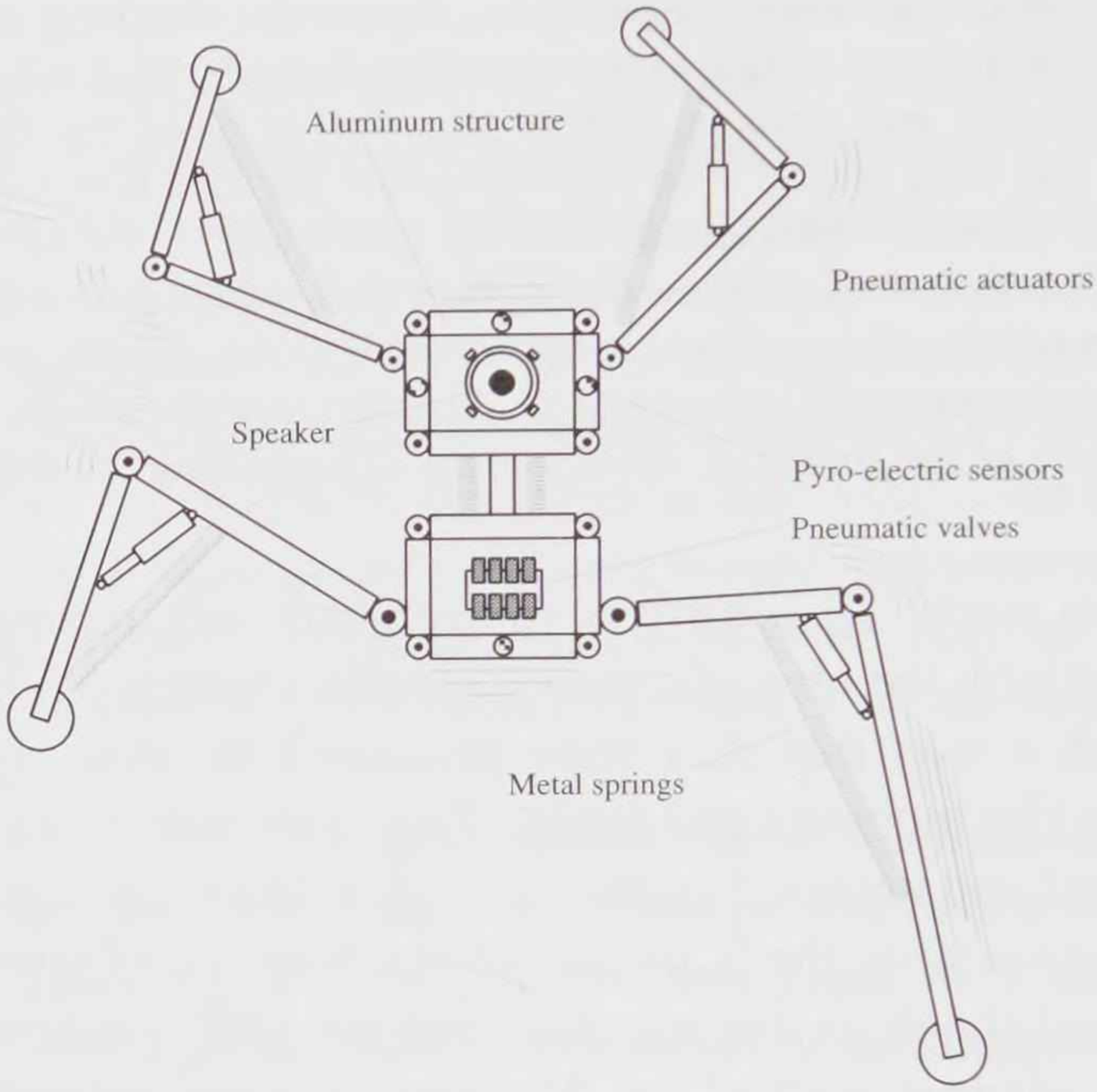


Figure 2. The Convulsive Machine (© L.P. Demers & B. Vorn 1997)

agents are created with specific features and parameters, and higher level emerging structures and behaviors are relatively foreseeable and unsurprising. For example, ant colony simulations are conceived to look and behave like real ant colonies, even if ant agents are still programmed individually on a lower level. Emergence is generally what we expect from these systems.

7. Truth, Trust, Believability and other Illusions

Automatons are still artifacts, as they do not reproduce, evolve and survive by themselves. They are phenotypes of our own memes, which do evolve. Artificial Life is the theater of a quest for the Double, it is an art of illusion that can be as convincing as other forms of representation such as cinematography. The semiotics of AL shows that most agents tend to mask themselves behind a biomorphic interface; they have names, visual features or behaviors that strongly suggest similarities with real living beings.

The illusion of intelligence is by no means necessary to the illusion of life, but the opposite is. For example, many chatterbots (like MegaHal[5]) show

poor intelligence in conversation situations although they are lively and entertaining as interlocutors. Also, most of our robotic installations are programmed with behavior control software blocks that switch more or less randomly from one to another in a sort of wild subsumption structure (Brooks 1986) without any need for a complex intelligence algorithm. The result is still the illusion of living creatures; their intelligence level is mostly a matter of the viewer's own subjectivity and projection.

8. Conclusion

Agents, bots, computer viruses, Tamagotchis and robots, these are names for different cybernetic organisms that have at least one thing in common: they are basically vehicles of action (agents) and meaning (media) between human beings. As automatons, virtual characters, synthetic actors, intelligent agents and biomorphic robots are invading an ever-greater part of our technocultural life, it is instructive to discuss their implication as virtual links between people.

The preliminary results of the survey show that there is still some reluctance from the users to consider AL as a media, although most people agree that it does produce signification and meaning.

But unlike traditional forms of expression such as film or theater, rigid scripts and stage plays have to be given up in favor of individual behavior design and uncertainty in the unfolding of events. In order for AL to attain recognition as an authentic means of communication, creating believable personalities for agents is an important goal that should combine the art of impressing human senses and the art of programming machinic behaviors.

9. Future Work

The present results are still in very raw form and more details will certainly emerge as further analysis is performed on the data from the questionnaires. Individual interviews with users may also unveil unexpected points of view on the interpretation of characteristics of AL agents (visual aspect, behavior, etc.). AL and its manifestations are continuously evolving. New surveys will probably be made, focusing on more particular categories and on more spe-

cific features of these categories. In this eventuality, creators and users should also be asked to fill out distinct questionnaires in order to clearly differentiate production from reception strategies.

As an artist working with robotics, my current work includes the creation of a multimedia performance staging a world populated exclusively by robot actors. This live project will present a machinic universe of automatons and cybernetic organisms expressing metaphoric behaviors, a surrealistic immersive environment where spectators are both visitors and intruders. Unlike our previous work, we do not intend to create a single exclusive installation, but rather a stage performance specifically defined in time and space and immediate in its relation to the public. This performance will be a symbolic expression of the trial of machines by men, as well as the trial of men by machines. It will act as a metaphorical and reflexive tribunal where identities intermix, where judges, jurors, victims and accused, take flesh in metal creatures born from our own conception of the world, of what is good and what is bad, of what is alive and what is not.

Acknowledgements

I would like to thank the Laboratoire de technologies interactives of the Communication department of UQAM (Université du Québec à Montréal) for hosting the survey site on its web server. Thanks also to Marc Lavallée for writing the Perl script for the on-line questionnaires. Our robotic installation projects are funded by the Canada Council for the Arts and the Conseil des arts et des lettres du Québec.

Notes

1. Demers, L.P. and B. Vorn, *Robotic Art Infopage*, www site on robotic and interactive installations, http://www.comm.uqam.ca/~vorn/chaos.html
2. http://www.comm.uqam.ca/~vorn/questionnaire/ALIFEsurvey.html
3. These properties are
   - a pattern in spacetime
   - self-reproduction
   - information storage of a self-representation
   - metabolism
   - functional interactions with the environment

– interdependence of the parts

– stability under perturbations

– the ability to evolve

– the ability to grow

4. Neumark, N., *Broadcast on Arts Today*, ABC Radio National (Australia), Oct. 16, 1997

5. Hutchens, J., *MegaHal*, http://www. ciips.ee.uwa.edu.au/~hutch/hal/HAL/

References

Brooks, R. 1986. A Robust Layered Control System for a Mobile Robot. IEEE *Journal of Robotics and Automation*: 14–23.

Cohen, J. 1968. *Les robots humains dans le mythe et dans la science*. Paris: Librairie Philosophique J. Vrin.

Demers, L.P. & B. Vorn 1995. Real Artificial Life as an Immersive Media. *Convergence: 5th Biennial Symposium for Arts and Technology proceedings*. New London, Conn.: 190–203.

Demers, L.P. & B. Vorn 1998. *Schizoid Ontologies of Cybernetic Lures*: to be published.

Doyne Farmer, J. & A. d'A. Belin 1992. *Artificial Life: the Coming Evolution*. In Artificial Life II: Addison Wesley: 815–833.

Franklin, S., and A. Graesser. 1997. Is it an Agent, or just a Program? A Taxonomy for Autonomous Agents. *Intelligent Agents III*: Springer-Verlag: 21–35.

Gilbert, D., et als. 1995. *The Role of Intelligent Agents in the Information Infrastructure.* IBM corp.

Gruber, T. 1993. A Translation Approach to Portable Ontologies. *Knowledge Acquisition* (2): 199–220

Kac, E. 1997. Foundation and Development of Robotic Art. *Art Journal*: 56(3): 60–67.

Leonard, A. 1997. *Bots: The Origin of New Species*. HardWired.

Levy, S. 1992. *Artificial Life: The Quest for a New Creation*. Pantheon.

McLuhan, M. 1968. *Pour comprendre les médias* (Understanding Media): Hurtubise HMH.

Penny, S. 1997. *Embodied Cultural Agents: at the Intersection of Art, Robotics and Cognitive Science*. Paper presented at the AAAI Socially Intelligent Agents Symposium. MIT.

Stubbs, B. & L. J. Hoffman 1990. *Rogue programs: viruses, worms, and Trojan horses.* George Washington University.

CHAPTER FIFTEEN

Agents as Artworks and Agent Design as Artistic Practice

Simon Penny

1. Introduction

My intention in this essay is to discuss agent building from the perspective of the visual arts. I will argue for the value of artistic methodologies to agent design. I will not advance some futuristic version of the romantic bohemian artist, agonising over an expressionistic agent in his garret. Nor will I propose the harnessing of artistic minds to the industrial machine. I want to advance another argument which is pertinent specifically to the building of Social Agents. I propose that there are aspects of artistic methodology which are highly pertinent to agent design, a which seem to offer a corrective for elision generated by the often hermetic culture of scientific research.

When one mentions the uses and functions of art in a scientific context, the understanding is often of superficial manipulation of visual 'aesthetic' characteristics in the pursuit of 'beauty' or a cool-looking demo. A more sophisticated approach recognises that the holistic and open ended experimental process of artistic practice allows for expansive inventive thinking, which can usefully be harnessed to technical problem solving (this has been the MIT Media Lab position). This approach tacitly recognises that certain types of artistic problem solving compensate for the 'tunnel vision' characteristic of certain types of scientific and technical practice.[1]

I have observed previously that the approach to the production of artworks by the scientifically trained tends to be markedly different from the approach of those trained in the visual arts. A case example is the comparison

of two works which I included in the Machine Culture exhibition at SIGGRAPH 93.[2] The *Edge of Intention* project by Joseph Bates and the Oz group at Carnegie Mellon University was an attempt to construct a knowledge base of plot structure and character development by distilling English literature and drama. Although the project had been in progress for several years, the developers admitted that it was still in its infancy. The audience experience at present was somewhat simplistic: the user (incarnated as one of the agents) could play childlike games (chasing and hiding, etc) with a group of cartoon entities which resembled moody jelly beans. The goal of the group was not to produce agents which were simulations of people, but which were 'believable' in their own terms. This 'believability' implies an abstraction of what we perceive to be intelligent behavior.

Luc Courchesne's *Family Portrait,* on the other hand, was comparatively low-tech. It consisted of four stations: four laserdiscs, four monitors and four Macintosh classics each with a simple HyperCard stack. The users stood and chatted with interactive video images. Although the interface consisted of using a trackball to choose responses to questions posed by the characters on the screen, the simulation of human interaction was uncanny. The artist has great finesse at simulating human interaction in the social space of the interface, a skill I have called "interactive dramaturgy". A particularly effective trick was that the four virtual characters would occasionally break their conversation with the visitors, and turn to interrupt or contradict each other. This illusion of 'real people' was aided by the handling of the hardware. The computer and AV hardware was hidden, even the monitor was hidden, the images were reflected in oblique sheets of glass in the darkened space, and seemed to float. Though low tech, *Family Portrait* was dramatically persuasive in a way that *Edge of Intention* was not.

The difference in approach of these projects illustrates my argument. One might generalise in this way (with apologies to both groups): artists will kluge together any kind of mess of technology behind the scenes because the coherence of the experience of the user is their first priority. Scientists wish for formal elegance at an abstract level and do not emphasise, or do not have the training to be conscious of inconsistencies in, the representational schemes of the interface. Arising from the tradition of Artificial Intelligence, the Edge of Intention project seeks to create general tools for an interactive literature by analysing the basic components of (rudimentary) social interactions, and building a system for their coordination. The focus of the effort was to build an

elegant and general internal system. The interface seemed to be a necessary but secondary aspect, like the experimental demonstration of a proof. The average user, however, will never gain access to the hermetic territory of the architecture of the code, and remains frustrated by the unsatisfying and incomplete nature of the representation of the system in the interface. Courchesne, on the other hand, does not attempt to build a general purpose system, but presents a seamless and persuasive experience for the user. Artists are trained to understand the subtle connotations of images, textures, materials, sounds, and the way various combinations of these might conjure meaning in the mind of the viewer. Artists must be concerned with the adequate communication of (often subtle) ideas through visual cues. They understand the complexity of images and the complexity of cultural context. Of course, the artistic solutions are often highly contingent and specific to a certain scenario, and may not generalise to general principles for a class of scenarios. This is not their goal.

While more academic disciplines valorise and reward a 'hands-off' approach, rewarding the more purely theoretic, artists are taught to integrate the artisanal and the conceptual (Penny 1997). Artistic practice is the shortest route to the physical manifestation of ideas. According to the traditional view, properly trained, the manual skill of the artist becomes an automatic conduit for the expression of abstract thought. Purely perceptuo-motor and abstract conceptual process are combined. Artists are judged on the perceived performance of a physically instantiated work, not on the coherence of a theory which may be demonstrated, perhaps obscurely. Criteria for a successful work is based almost solely on its influence on the viewer. An artwork must motivate the viewer to engage intellectually and emotionally with the work. In a good work, the 'interface' is finely honed and engagement should develop over the long term through.

This condition of engagement is a paradigmatic case of what Jonathan Crary calls the 'techniques of the observer' (Crary 1992). In the book of the same name, Crary argues that pictures would remain meaningless and mute without the unconscious and uncelebrated training of observers, as a cultural group. We are all trained in how we look at and appreciate pictures.[3] The meaning of a work is negotiated by the observer in the moment of looking. Meaning is construed entirely as a result of the observers' cultural training.

A salutary example of the cultural specificity of this training is the history of depiction of 'new lands' by colonising peoples. Take for instance the

depiction by the British colonists of Australia in the closing years of the 18th century and later. Almost invariably in these pictures, aboriginals look negroid, eucalypts look like elms, kangaroos look like giant pudgy mice and the Australian bush looks like rolling English countryside. It took over 100 years until painters captured the quality of the Australian light. This example demonstrates that what we see depends to a great extent on what we have been trained to see. We extrapolate from our previous experience to explain our new perceptions.

Over the past decade, my artistic practice has developed from the construction of sensor driven interactive installations to systems with at least rudimentary forms of agency. My focus of interest has been for several years what I call the 'aesthetic of behavior', a new aesthetic field opened up by the possibility of cultural interaction with machine systems. I have the luxury of being able to experiment with the modalities of systems, without being constrained by an externally specified task for the system. A secondary interest arising from the first is the potential application of various 'Alife' techniques as artistic tools, producing artworks which demonstrate behaviors which go beyond a 'locked-down' state machine model. This combination of interests leads me inevitably into agent design. My background in art predisposes me to integrated, holistic, situated and embodied practice (both by the maker and in the agent).

In my own practice I tend to define the envelope of the problem first: the system has to do this on these occasions in this way, it has these physical constraints, this power limitation, etc. From these specification I work slowly inward from desired behavioral to physical structure to specifics of sensing and actuation, often specifying hardware first, eventually arriving at the set of constraints within which the code must function. Contrarily, computer scientists have a tendency to look briefly at the surface level, identify a 'problem' that might respond to a rule-based solution, then dive deep into the abstractions of code at the most conceptual level, building the ramifications of a conceptual design up through the more abstract to the more 'mechanical' aspects of the code, finally surfacing to look back at the interface and see if it works. This approach results in fragmentary and inconsistent interfaces.

These are some of the values which I bring into my robotic and agent practice. These positions bring me close to many already established in Cybernetics and in critiques of traditional AI which concern themselves with groundedness, embodiment, situated cognition and emergent behavior, as

discussed by Agre, Brooks, Cariani, Dreyfus, Johnson, Varela, et al. (Agre 1997; Brooks 1991; Dreyfus 1992; Johnson 1987; Varela, Thompson and Rosch 1993) By the same token, my training steers me away from the sensibilities of symbolic AI approaches.[4] In the following text I will discuss three recent works as examples of the way these positions arise or are applied.

2. Petit Mal

The goal of the project *Petit Mal: An autonomous robotic artwork* was to produce a robotic artwork which was truly autonomous; which was nimble and had 'charm'; that sensed and explored architectural space and that pursued and reacted to people; that gave the impression of intelligence and had behavior which was neither anthropomorphic nor zoomorphic, but which was unique to its physical and electronic nature (see Plates 1 and 2). Petit Mal was conceived in 1989, construction began in 1992. Since its public debut in February 1995 it has proven to be reliable and robust, it has been shown in many festivals where it must interact with the public continuously for 8 hour days, for weeks at a time.

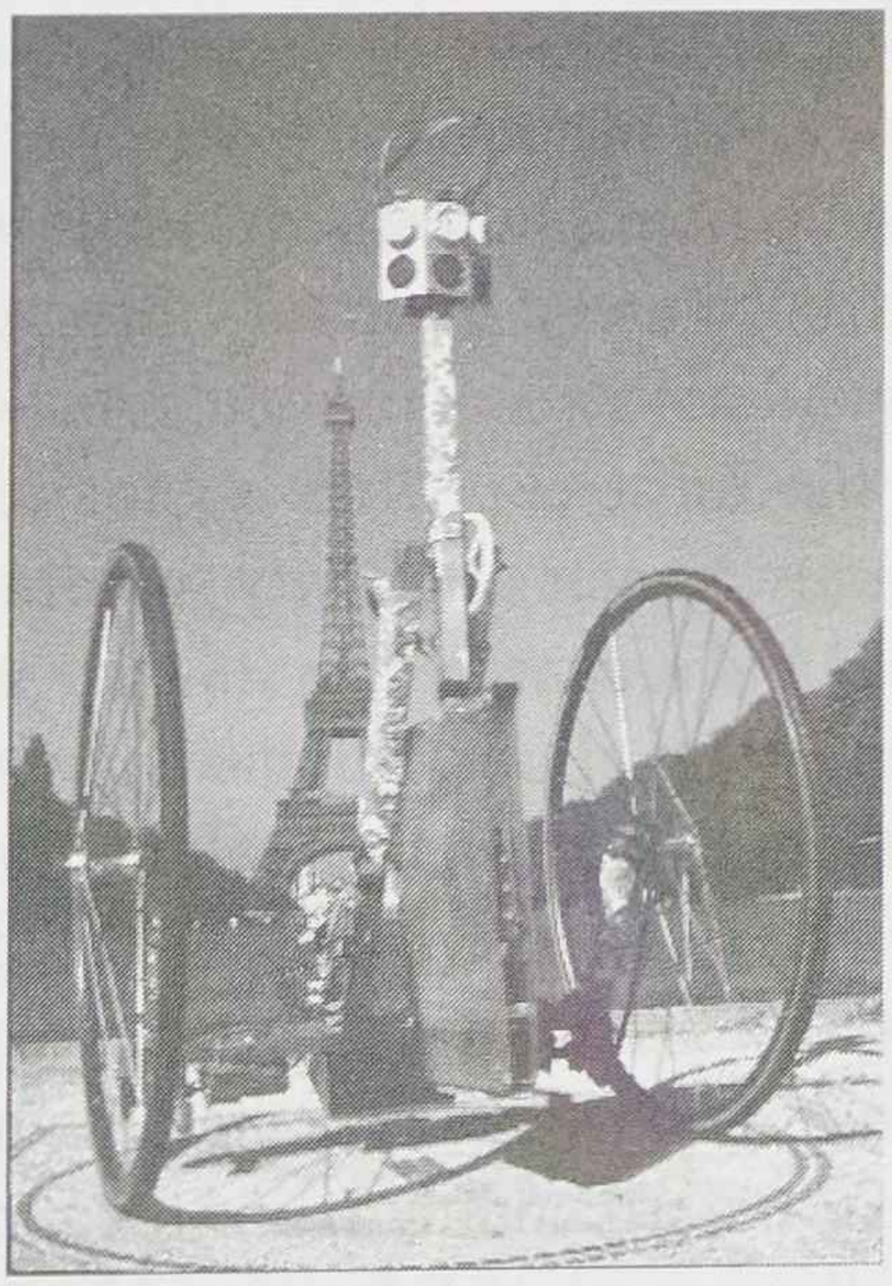


Plate 1. Petit Mal: An Autonomous Robotic Artwork

It was not my intention to build an artificially intelligent device, but to build a device which gave the impression of being sentient, while employing the absolute minimum of mechanical hardware, sensors, code and computational power. The research emerged from artistic practice and was thus concerned with subtle and evocative modes of communication rather than pragmatic goal based functions. My focus was on the robot as an actor in social space. Although much work has been done in the field of screen-based interactive art, the 'bandwidth' of interaction in these works is confined by the limitations of the desktop computer. I am particularly interested in interaction which takes place in the space of the body, in which kinesthetic intelligences, rather than 'literary-imagistic' intelligences play a major part. I conceive of artistic interaction as an ongoing conversation between system and user rather than the conventional (Pavlovian) stimulus and response model.

Acknowledging that there is no canon of autonomous interactive esthetics, *Petit Mal* is an attempt to explore the aesthetic of machine behavior and interactive behavior in a real world setting. Every attempt was made to avoid anthropomorphism, zoomorphism or biomorphism. It seemed all too easy to imply sentience by capitalising on the suggestive potential of biomorphic elements. I did not want this 'free ride' on the experience of the viewer. I wanted to present the viewer with a phenomenon which was clearly sentient, while also being itself, a machine, not masquerading as a dog or a president.

I wanted to build a device whose physiognomy was determined by brutally expedient exploitation of minimal hardware. The basic requirements of navigation and interaction with humans determined the choice of sensors. The suite of sensors is absolutely minimal: three ultrasonics, three pryo-electrics, two very low resolution encoders and a low-tech accelerometer. The dicycle design offered the most expedient motor realisation for drive and steering but demanded a low center of gravity to ensure stability. This swinging counterweight would have caused the sensors swing radically, looking first at the ceiling then at the floor, so the sensors were mounted on a (passively stabilising) second internal pendulum. In this way the structure specified the necessary extrapolations to itself, the development of the mechanical structure was not a gratuitous design but a highly constrained and rigorous engineering elaboration based on the first premise of two wheeled locomotion. The lower or outer pendulum carries motors, motor battery and motor drive electronics, the inner pendulum carries the sensors at the top and processor and power supplies as counterweight in the lower part. The batteries

are not dead weight but in both cases also function as the major counter-weights. In an analogy to the semi-circular canals of the inner ear, an accelerometer at the pivot of the inner pendulum is a rudimentary proprioceptive sensor, it measures relationships between parts of the robot's 'body'. It was important to me that this robot was 'aware' of its body.

From the outset I wanted to approach hardware and software, not as separate entities but as a whole. I wanted the software to 'emerge' from the hardware, from the bottom up, so to speak, The code would make maximal utilisation of minimal sensor data input. *Petit Mal* has had four successive sets of code, each increasingly more subtle in its adaptation to the dynamics of the device and more effectively exploiting the minimal processor power (one 68hc11). My approach has been that a cheap solution (in labor, money or time) to a particular problem which was 70% reliable was preferable to a solution which was 90% reliable but cost several times as much. It was pointed out to me by an engineer that my 'under-engineering' approach could lead to a much wider range of possible (though unreliable) solutions. The field of possibility is thereby expanded. Eventually such solutions could be refined. He was of the opinion that this approach could lead to better engineering solutions than an approach which was hindered by a requirement of reliability in the research phase.

In robotics circles one hears the expression 'fix it in software' applied to situations when the hardware is malfunctioning or limited. This expression is emblematic of a basic precept of computer science and robotics, the separation of hardware an software and the privileging of abstract over concrete. I attempted, in *Petit Mal*, an alternative to this dualistic structure. I believe that a significant amount of the 'information' of which the 'intelligence' of the robot is constructed resides in the physical body of the robot and its interaction with the world.

A 'Petit Mal' is an epileptic condition, a short lapse of consciousness. The name was chosen to reflect the robot's extremely reactive nature, *Petit Mal* has essentially no memory and lives 'in the moment'. My approach has been that the limitations and quirks of the mechanical structure and the sensors are not problems to be overcome, but generators of variety: the very fallibility of the system would generate unpredictability. My experience has shown that 'optimization' of the robots behavior results in a decrease in the behaviors which (to an audience) confer upon the device 'personality'. In sense then, my device is 'anti-optimised' in order to induce the maximum of personality. Nor

is it a simple task to build a machine which malfunctions reliably, which teeters on the threshold between functioning and non-functioning. This is as exacting an engineering task as building a machine whose efficiency is maximised.

2.1 *Behavior, interaction, agency*

The example of Australian colonial painting (cited above) is pertinent in the explanation of peoples' behavior toward *Petit Mal* and the way it will change. Almost invariably, people ascribe vastly complex motivations and understandings upon *Petit Mal*, which it does not possess. Viewers (necessarily) interpret the behavior of the robot in terms of their own life experience. In order to understand it, they bring to it their experience of dogs, cats, babies and other mobile interacting entities. In one case, an older woman was seen dancing tango steps with it. This observation emphasises the culturally situ-



Plate 2. Unidentified visitor interacting with Petit Mal, Voyage Virtuel exhibition Paris 1995

ated nature of the interaction. The vast amount of what is construed to be the 'knowledge of the robot' is in fact located in the cultural environment, is projected upon the robot by the viewer and is in no way contained in the robot. The clear inference here is that, in practical application, an agent is first and foremost, a cultural artifact, and its meaning is developed, in large part, by the user and is dependent on their previous training. This means that, in the final analysis, an agent is a cultural actor, and building an agent is a cultural act. Here the rarefied and closed proof system of science is ineffably forced into engagement with the world.

Such observations, I believe, have deep ramifications for the building of agents. Firstly, any effective agent interface design project must be concerned with capitalising on the users' store of metaphors and associations. Agents work only because they trigger associations in the user. So agent design must include the development of highly efficient triggers for certain desired human responses. In his painting *Ceci n'est pas un pipe*, René Magritte encapsulated the doubleness of symbols and the complexity of representation. This doubleness can be used to good effect in agent design: a very simple line drawing (of a pipe, for instance) triggers a rich set of associations in the user. However, for the same reasons, these associations, like any interface, are neither universal nor intuitive, they are culturally and contextually specific.

Another curious quality of *Petit Mal* is that it trains the user, due to their desire of the user to interact, to play; no tutorial, no user manual is necessary. People readily adopt a certain gait, a certain pace, in order to elicit responses from the robot. Also, unlike most computer-based machines, *Petit Mal* induces sociality amongst people. When groups interact with *Petit Mal,* the dynamics of the group are enlivened. Readers from the agent research area might wonder at this point if the systems I describe might be appropriate for various sorts of application domains. I would respond: 'probably not', nor is this my goal. I am interested in the modalities of interactive systems as new cultural environments. And I would reiterate my argument that because I am able to experiment without the constraint of total reliability or a pragmatic work-oriented goal, I can open up a wide field of possibilities, some of these possibilites may ultimately have application or relevance in pragmatic applications.

3. Sympathetic Sentience

Sympathetic Sentience is an interactive sound installation which generates complex patterns of rhythmic sound through the phenomenon of 'emergent complexity'. *Sympathetic Sentience* is an attempt to build a physically real model of emergent complex behavior amongst independent units, which produces constantly changing sound patterns. As with *Petit Mal*, there was an interest in designing the most technologically minimal solution, in this case, for a system which would demonstrate persuasively 'emergent' behavior.

Each of the 12 comparatively simple, identical electronic units alone is capable of only one chirp each minute. Rhythmic and melodic complexity develops through a chain of communication among the units. In the installation, each unit passes its rhythm to the next via infrared signal. Each unit then combines its own rhythm with the data stream it receives, and passes the resulting new rhythm along. Thus the rhythms and timbral variations slowly cycle around the group, increasing in complexity. The system is self-governing, after an initial build-up period, the system is never silent nor is it ever fully saturated.

The 12 units are mounted on the ceiling and walls of a darkened room. The experience of the visitor is of an active sound environment of 12 'channels' in which there is recognisable, but not predictable, patterning. The visitor can interrupt this chain of communication by moving through the space. This results in a suppression of communication activity and hence reduction of complexity. A long interruption results in complete silencing of the whole group. When the interruption is removed, slowly a new rhythm will build up. The build-up of a new rhythm cycle can take several minutes. The rhythm cycles are never constant but continually in development. To gain a sense of the full complexity of the piece, it is necessary to spend several minutes with the piece in an uninterrupted state.

3.1 *Technical realisation*

Jamieson Schulte and I have built several iterations of the work. *Sympathetic Sentience One* was built entirely in hardware logic (TTL ICs). The basic premise is extremely simple: each unit is receiving, processing and forwarding a continuous stream of data. Each unit 'edits' that stream 'on the fly', adding or omitting an occasional bit. This editing is done in such a way that the

'density' of the sound is 'self-governing'. The critical part of each unit is an exclusive OR gate. On each unit, the signal is received by an IR receiver, demodulated and sent to a shift-register (delay). Emerging from the delay it meets a feed from the on-board oscillator at the exclusive OR gate. The signal emerging from the gate goes to both the IR emitter and the audio amplification circuit. The units communicate in modulated infrared signals using hardware similar to that used in TV remote controls.

While in *Sympathetic Sentience One*, only the rhythmic patterns were subject to change through the emergent complex behavior, in *Sympathetic Sentience Two*, other sound characteristics such as pitch and envelope are also subject to gradual change through the emergent complex process. To achieve this, *Sympathetic Sentience Two* uses small microprocessors (PICs) to replace the hardware logic.

3.2 *Emergence*

Whether this behavior is deemed to be 'emergent' is a matter of previous experience. Most visitors find it reminiscent of the sound of communities of frogs, crickets or cicadas. But to at least one rather dry observer, it was simply a chaotic system of a certain numerical order. To another is was a demonstration of one model of neural propagation. Here emergence would seem to be 'in the eye of the beholder'.

The term 'emergence' seems to be defined rather loosely, even in scientific texts. In some cases it is applied to the interaction of two (or more) explicit processes which result in a third 'emergent' process which was, however, entirely intended. Similarly, the fitness landscape of Stuart Kauffman establish a desired end condition (Kauffman 1993). This would seem to be a rather different and narrower sense of emergence than that of the termite community, though attempts to reproduce such behavior in programmable models, such as the stigmergic multi-robot systems of Beckers, Holland and Deneubourg, reduce the complex interactions to deterministic events (Beckers, Holland & Deneubourg 1994). The paradigmatic 'emergent' systems are the development of the mind/brain and the process of genetic evolution. The difference here is that these systems are open ended, goal states are not specified.

4. Fugitive

Fugitive is a single user spatial interactive environment. The arena for interaction is a circular space about 10m dia. A video image travels around the walls in response to the users position (see Plate 3). This is the simplest level of interactive feedback: the movement of the image, tightly coupled to the movement of the user, is an instantaneous confirmation to the user that the system is indeed interactive. The behavior of the system is evasive, the image, in general, runs away from the user. The user pursues the image. Over time the response of *Fugitive* becomes increasingly subtle and complex (constrained by the need to be 'auto-pedagogic', to continually more or less make sense to the user). A user must spend almost 15 minutes to get through the full seven chapters and elicit the most complex system responses.

The user is totally unencumbered by any tracker hardware, sensing is done via machine vision using infra-red video.[5] The space is lit with 13 infra-red floodlights. User tracking is achieved via a monochromatic video camera mounted vertically upwards, looking into a semi-circular mirror suspended in the center of the room. Preliminary vision processing occurs on a PC. Two streams of serial data are output. Simple angular position data is sent to the custom PID motor control board to drive the projector rotation motor. Values for MAE calculations are sent to the MAE2 (Mood Analysis Engine2) running on an SGI 02 computer. On the basis of this calculation, the VSE (Video Selector Engine) selects, loads and replaces digital video on a frame by frame basis. Video data is fed to the video projector.

The user is engaged in a complex interaction with the system. The basic logic of interactive representation in *Fugitive* amounts to this: user movement is represented by camera movement within the image, and image movement across the wall. The segwaying of image content and its physical location is the 'expression' of the system. The output of the Mood Analysis Engine controls the flow of digitised video imagery in such a way that no two people walking the same path in the installation will produce the same video sequence, because their bodily dynamics are different. The system responds to the dynamics of user behavior and their transitions over time. Ideally, the system responds not simply to changes in raw acceleration or velocity or position, but to kinesthetically meaningful but computationally complex parameters like directedness, wandering or hesitancy. This is achieved in a multi-stage process of computationally building up the complexity of param-

eters. The input level data from the vision system is limited to raw position in each frame. From this, simple values for velocity and acceleration are calculated. A third level of more complex parameters is then constructed: average acceleration over various time frames, variance and so on. Finally, values for various combinations of these parameters are used to determine the entry and exit points for 'behaviors' which are matched to video selections.

The images do connect with some small degree of semantic significance, there is a minimal hypernarrative, but characterisation and plot structure were explicitly avoided. The chosen imagery is lanscape, each 'chapter' being a specific location at a specific time of day. An hypertextual structure and a logic of transition links one 'chapter' or location with the next. As time progresses, the user propels themselves through seven location chapters. A formal garden sequence is a kind of 'vestibule'. You go there at the beginning and return there between each chapter. When you got the center the projector slowly rotates and shows you a series of archways. You choose to set out of the center (metaphorically through one of the archways) and you make the

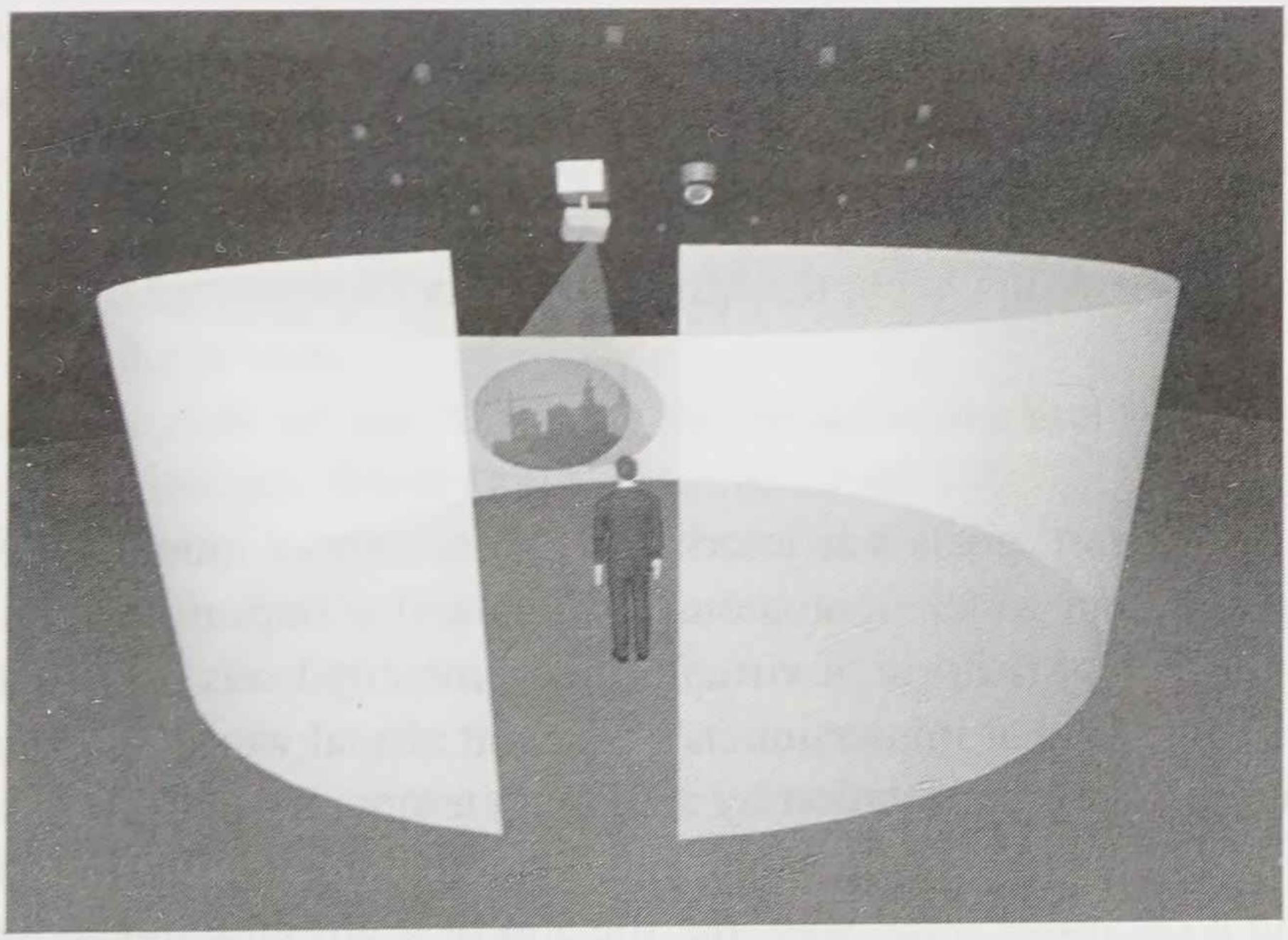


Plate 3. Schematic rendering of Fugitive showing circular screen/room with entrance, video image projected from motion controlled video projector in center of ceiling, semi-circular mirror used by vision system, ceiling mounted infra red floodlights. (Rendering by Armen Oulikhanian, ZKM 1997)

transition into a new chapter. This is the only case in which particular imagery is connected with a specific location in the room. When you have explored the chapter adequately (as determined by the system), you transition back into the 'garden'. All other video material is located 'temporally' and triggered dynamically rather than positionally. This reinforces the continuity of body and time, against the continuity of an illusory virtual space.

The system is completely free of textual, iconic, or mouse/buttons/menus type interaction.

In building *Fugitive*, my concern was with the aesthetic of spatial interactivity, a field which I regard as being minimally researched. Watching spatial interactives over several years, I was frustrated by the simplistic nature of interaction schemes based on raw instantaneous position and simple state-machine logic. I wanted to produce a mode of interactivity which did not require the user to submit to a static Cartesian division of space (or simply groundplane). I wanted to make an interactive space in which the user could interact with a system which 'spoke the language of the body', and which critiqued VR and HCI paradigms by insisting on the centrality of embodiment. I wanted to develop a set of parameters which could be computationally implemented, which truly reflected the kinesthetic feeling of the user, their sense of their embodiment over time. *Fugitive* is an attempt to build an entirely bodily interactive system which interprets the ongoing dynamics of the users body through time as an expression of mood. I called this part of the code (somewhat tongue-in-cheek) the Mood Analysis Engine.

4.1 *Immersion and embodiment*

One of my 'covert' goals was to critique the rhetoric of immersion in VR by building a system which continuously offers and collapses such an illusion. The last decade of rhetoric of virtualisation probably leads users to expect or hope for some kind of immersion in a coherent virtual world. *Fugitive* explicitly contradicts this expectation by setting up temporary periods in which the illusion of immersion is believable, and then breaking the illusion. If the user moves in a circumferential way, the illusion of a virtual window on a larger world is created. As you move, say, to the left around the perimeter, you will see a pan as a moving 'virtual window'. As you continue it will segway into another pan. If you reverse your direction, the same pan will occur in reverse, but when you get to the beginning of pan2, you segway to pan3, not pan1. In

this way the illusion of a virtual world seen through a virtual window, is collapsed.[6]

In conventional systems, the illusion of immersion is positional, the absolute position of the tracker (etc) corresponds to a specific location in the virtual world. Such a virtual world, a machinic system, maintains a rather repressive continuity: the continuity of the illusory architectural space. In *Fugitive*, the continuity of the system is a phenomenological one focused on the continuity of embodiment, not the instrumental one of a consistent virtual space in which the body is reduced to little but a pointer. *Fugitive* is not positional, the primary and structuring continuity is the deeply subjective continuity of embodied being through time.

4.2 *Embodied looking: Imagery as the voice of the agent*

Fugitive is about the act of looking, embodied looking, and it is about the metaphorisation of looking via video. The title '*Fugitive*' emphasises the evanescence of the experience of embodied looking. The attempt is, rather perversely, to avoid eliciting the kind of externalised interest in imagery and subject matter which one has when looking at a painting. This is because to goal is always to fold the attention of the user back onto their own sense of embodiment and the functioning of the system in relation to their behavior. *Fugitive* is not primarily a device for looking at pictures (or video), it is not a pictorial hyper-narrative. It is a behaving system in which the video stream is the 'voice' of the system.

I want the user to see 'through' the images, not to look only at the 'surface' of the images. Strictly speaking, this meant I should choose imagery that was inherently uninteresting. The exercise is of course fraught with paradox, especially for the scopically-fixated viewer. The user is presented with a darkened circular space the only changing feature of which is a changing image, and yet the user is encouraged to understand the image primarily as an indicator of the response of an otherwise invisible system.

4.3 *The auto-pedagogic interface*

An interactive work is a machine, and one must learn to operate a machine. But visitors to artworks are seldom previously trained. Although prior training has become a part of theme park amusements, nobody wants to do a tutorial or

read a manual before they experience an artwork. Nor do I find it acceptable for the user to have to don 'scuba gear' (to borrow Krueger's term) before entering the work. A user should be able to enter unencumbered by special clothing or hardware. So a central issue in interactive art is managing the learning curve of the user. One solution is to make a work is so simple in the dynamics of interaction that it is easy to understand but immediately boring. Alternatively, works can be so complex that the average user cannot discern the way in which they are controlling or effecting the events, it appears random. In avoiding these two undesirables, the artist must either choose a well known paradigm (such as monitor-mouse-buttons or automobile controls) or if one desires the modalities of an interface which is novel, then the user must be trained or the system must teach the user.

I cannot endorse the concept of the 'intuitive' interface because it implies a naive universalism and an ignorance of cultural specificity, aspects of which I noted in my discussions of 'techniques of the user' and colonial painting. In *Petit Mal* I discovered that if the user is propelled by a desire to interact, that learning will occur in an unimpeded and transparent way. In *Fugitive*, I attempted to formally produce this effect in a much more complex system. Such an 'auto-pedagogic' interface must present itself as facile to a new user, but progressively and imperceptibly increases in complexity as the familiarity of the user increases. Transitions to higher complexity should be driven by indicators of the behavior of the user.

In the current implementation of Fugitive, in order to ensure that the 'interface' be 'auto-pedagogic', the system exhibits only two behaviors at the beginning. Others are introduced along the way, and control of transitions becomes more complex. In future implementations, system behavior will be more 'intelligent', as an agent which learns and expresses certain 'desires'.

4.4 *Poetics of interaction*

The degree to which the changes in output are interpreted by the user as related to their behavior is a key measure of the success of any interactive system. Ideally, changes in the behavior of the system will elicit changes in the users behavior, and so an ongoing 'conversation' rather than a chain of 'Pavlovian' responses will emerge. Art artwork is by definition not literal or didactic, it is concerned with poetic and metaphoric associations. So an interactive artwork should not simply tell you something like 'you have mail'. Nor would it be

interesting if *Fugitive* told you: 'you just moved two paces left'. The goal is to establish a metaphorical interactive order where the user's movement 'corresponds' to some permutation of the output. It is all too easy to produce a system which the user cannot distinguish from random behavior. The designer must successfully communicate that the user is having a controlling effect on the system and at the same time engage the ongoing interest of the user with enough mystery. One hopes for some poetic richness which is clear enough to orient the user but unclear enough to allow the generation of mystery and inquisitiveness. The system must engage the user, the user must desire to continue to explore the work. This is a basic requirement of any artwork.

4.5 *The paradox of interaction*

Representation of the response of the system back to the user is key to any interaction. Not only must one reduce human behavior to algorithmic functions, but one must be able to present to the user a response which can be meaningfully understood as relating to their current behavior. One can collect enormous sets of subtle data, and interpret it in complex ways, but if it cannot be represented back to the user in an understandable way, it is ultimately useless.

Having collected complex data with multiple variables, how do you build a rule-based system which establishes such fluid correspondences when the data base is a finite body of fixed video clips? The impossibility of this task was resoundingly brought home to me while making *Fugitive*. In the case of *Fugitive*, the sophistication of the response of the system had to be scaled back to a point where it could be represented by the video, the limitations of the rule based system which organises those clips into classes and the range of likely or possible behaviors in that circular geometry.

But images are complex things. Many types of information can be extracted from a single still image, let alone a moving image sequence. A major difficulty in the interactive scheme of *Fugitive* is for the user to determine which aspects of the images presented signify the expression of the system. Is the presence of red significant, the presence of water or a tree? Is it a question of the direction of movement of various objects in the image or the quality of the light? In *Fugitive*; subject matter, color etc, do not carry meaning about the state of the system. The aspect of the image which is the 'voice' of the system is camera movement.

5. Conclusion

An artwork, in my analysis, does not didactically supply information, it invites the public to consider a range of possibilities, it encourages independent thinking. So building an interactive artwork requires more subtle interaction design than does a system whose output is entirely pragmatic, such as a bank automat. My work over the past decade has focused upon: the aesthetic design of the user experience given the diversity of cultural backgrounds and thus of possible interpretations; the development of embodied interaction with systems where the visitor is unencumbered by tracking hardware; the development of paradigms of interaction which go beyond state machine models to embrace and exploit Alife, emergence and social agent models. There is some divergence in current definitions of 'autonomous agents', more in the term 'socially intelligent agents'. While the works I have discussed are only marginally agents in Cariani's definition: "self-constructing, conscious, autonomous agents capable of open-ended learning", they do demonstrate a rich and complex interaction with the user.

I have emphasised the relevance of artistic methodologies to the design of social agent systems. Typically, artistic practice embraces an open ended experimental process which allows for expansive inventive thinking. Artistic practice emphasises the cultural specificity of any representational act, acknowledging that meaning is established in the cultural environment of the interaction, not in the lab. It emphasises the embodied experience of the user. And it emphasises the critical importance of the 'interface', because the interface of the agent, like an artwork, is where communication finally succeeds or fails.

Notes

1. This is not to denigrate the powerful influence that science and the scientific method have had over all fields of intelligent thought over the last century, art included.

2. I curated the Machine Culture exhibition of interactive art at SIGGRAPH93. (Penny, S. 1993)

3. As any parent has noticed, this happens very early in life. I noted with surprise that at 15 months, my son, who could yet speak only a few single words, already had a rich and precise understanding of visual representation at various levels of abstraction.

4. ...Pask's electrochemical assemblage has the distinct advantage of concreteness; it grounds us in the realm of the sensuously apprehendable material world. It is too easy to lose the broader research program, the *realisation* of self-constructing, conscious, autonomous agents capable of open-ended learning, when one descends into the infinite labyrinth of formal *descriptions* of how they might work. (Cariani, P. 1993)

5. Design for Fugitive was begun in May 1995. Construction began in May 1996. MAE and projector machine were built at Carnegie Mellon University, Pittsburgh PA, USA by Simon Penny and Jamieson Schulte. Digital video editing, MAE2 (Mood Analysis Engine2), VSE (Video Selector Engine) and full scale installation were built up at and the Institut für Bildmedien, Zentrum Für Kunst und Medientechnologie, Karlsruhe, Germany, March-May 1997, by Simon Penny and Andre Bernhardt.

6. The circularity of the space restructures all shots. Pan shots are laid over a certain specified arc in degrees/frame. The shot is 'unwrapped', the original landscape is thus compressed or stretched: a 90 degrees pan might be stretched over 180 degrees.

References

Agre, P. 1997. *Computation and Human Experience.* Cambridge University Press. Cambridge, New York.

Beckers, R., O. Holland, & J. Deneubourg. 1994. From Local Actions to Global Tasks: Stigmergy and Collective Robotics. In R. A. Brooks & P. Maes (Eds.), *Proc. Fourth International Workshop on the Synthesis and Simulation of Living Systems, Artificial Life IV*, MIT Press, 181–189.

Brooks, R. 1991. *Intelligence Without Reason* AI Memo #1293, (April) MIT.

Cariani, P. 1993. To Evolve an Ear: Epistemological Implications of Gordon Pask's Electrochemical Devices. *Systems Research* 10(3), 19–33.

Crary, J. 1992. *Techniques of the Observer: On Vision and Modernity in the 19th Century.* MIT Press.

Dreyfus, H. 1992. *What computers still can't do.* MIT Press.

Johnson, M. 1987. *The Body in the Mind.* University of Chicago Press.

Kauffman, S. 1993. *The origins of order.* Oxford University Press.

Penny, S. 1993. Machine Culture. In *ACM Computer Graphics SIGGRAPH93 Visual Proceedings Special Issue*, 109–184.

Penny, S. 1997. The Virtualisation of Artistic Practice: Body Knowledge and the Engineering Worldview. *CAA Art Journal 56(3)*, Fall 1997, Guest Editor Johanna Drucker, 30–38.

Varela, F. Thompson, E. and Rosch, E. 1993. *Embodied Mind.* MIT Press.

CHAPTER SIXTEEN

Living with Socially Intelligent Agents

A Cognitive Technology View[1]

Kerstin Dautenhahn and Chrystopher L. Nehaniv

1. Cognitive and Social Technology

Throughout human technological history, new tools and technologies such as agriculture, writing, books, telephones, radio, television, the Internet and e-mail, have had a profound influence on our relationships to ourselves, to others and to the world around us. Using a hammer provides us with a way in which to easily direct force for driving nails into wood, but use of such a tool can change the hand that wields it: Calluses may form on the fingers, a thumb may occasionally be struck accidently, muscles in the arm that control it build and adapt in certain ways in response to long use. Seeing two boards that need joining we seek out and employ our hammer, the hammer becomes an extension of the man. In the sense that our physical and cognitive tools (eyeglasses, hammers, automobiles, e-mail, numerical notations, etc.) are now just parts of our selves, we are already cyborgs (Haraway 1991; Nehaniv 1997) with various special abilities and cognitive calluses.

In this short chapter we are concerned with raising issues of optimizing the *cognitive fit* of humans and their new tools, agents, the impacts of agent technology on human social interaction and on human cognition, the changes in who or what we are as we come to rely on the power of agent technology. In Jacob Mey's words (Mey 1997), we are concerned with the fitting of the 'cyb' to the 'org'. We have little choice about the fact of our being cyborgs, of our using technology or being subject to technological impact, so in order to make

the best of things it makes sense to consider how our technology can be designed in the most humane ways. Complementary to e.g. Human-Computer Interaction (HCI) research, work in Cognitive Technology (CT) serves the interest of (individual) people, aims at improving human cognitive capacity by means of transcending human cognitive closures (via cognitive prosthetics), investigates ways in which people adapt to the demands of technology, and studies the holistic integration between human mental processes and technological progress (Gorayska, Marsh, and Mey 1997). In this chapter we give a Cognitive Technology view on social agent technology.[2]

The manner in which agent technology develops results (in part) from the *externalization* of human cognition (Gorayska and Marsh 1996; Gorayska and Mey 1996). Human designers create agents that they feel will fill some need, be useful as products, or be commercially successful. Yet as with all human tools, our use of technology changes who we are and how we think, and changes the environments we live in. This feeds back into human cognition and societies making further technological externalizations, i.e. new tools, conceivable, designable, desirable and possible. We are now creating or living with non-human agents who represent us in negotiations and transactions, speak for us on the telephone, arrange our schedules, select what information to show us from a mass of relatively unstructured information, filter our correspondence, assist us in learning a new piece of software, act as our secretaries, entertain us in films and games or serve us in other ways. This is of course just a partial list of a few areas where agent technology is already being applied today, and the previous chapters in this book give examples of research in these areas. As a consequence of agent technology, our environment is changing and ways of interacting that we may have only employed previously when interacting with other humans are forming the basis for interfacing with our agents.

Questions for the long-term cognitive and social effects of social agent technology we should ask ourselves are:

- How will agent technology affect human cognition?
- How will using or interacting with this technology change who we are?

If we do not ask such questions, the answers that come to us by letting things run their course in the future may be neither humane nor desirable. Experiencing relatedness to agents that one comes to think of as social beings, as surrogate humans, or, in effect, as substitutes, representatives and alter-egos

for our selves in our work and play, may affect how we interact with other humans, how we present ourselves to them, and how we think about them and link to them. Living with socially intelligent agents will transform us and our societies in important ways. It therefore makes sense for us as designers and users of agents, especially of agents that interact with humans as new partners in their social lives, to ask ourselves such questions. We need to demand and produce designs that are informed by considerations which seek to optimize the cognitive fit between humans as social beings and their new socially active and interactive tools, agents.

2. Costs and Rewards of Agent Technology for Humans

Some central cognitive technology questions regarding design trade-offs identified by Barbara Gorayska, Jonathon Marsh, and Jacob Mey (Gorayska, Marsh, and Mey 1997) for the creation and use of tools in general are adapted here to social agents. These questions make sense at a global level, i.e. at the level of global or culture-wide societal and psychological impacts of agent technology. Moreover and perhaps more usefully, one can ask these same questions at the level of each particular proposed application.

- How will it serve the interests of people?
- How will it enhance human cognitive or social capability?
- What will be the technologically induced human adaptations?
- Does it or how does it seek to integrate technological and human processes while respecting and preserving human wholeness?

Design choices involve many trade-offs. Each application of technology has its costs and rewards, which are likely to differ for different users, different applications and for different communities.

Some dimensions in which to ask such cost and reward questions for living with agent technology (in general) and for any particular instance of socially intelligent agents are sketched in the following list. Many of these dimensions and others may be discussed also in regard to interaction, social intelligence and a constructive perspective (Dautenhahn 1997, 1998, 1999). The members of this list can be considered realms of influence on the human cybernetic organism (Gorayska, Marsh, and Mey 1997).

Impact on Human Cognition and Social Interaction – Realms of Influence:

1. The realm of the individual:
   - Self, identity, personality
   - Autobiography and autobiographical development
   - Situatedness in time and space
   - Places providing meaning and context
   - Knowledge, skills, processing speeds
2. The realm of the individual in social interaction:
   - Believability
   - Social embeddedness
   - Social efficiency and appropriateness of responses in interaction and communication
3. The realm of society:
   - Emergence of communities and cultures
   - Norms, conventions, values
   - Memetic development

On all three levels (1.–3.) adaptation or better co-adaption plays an important role: individuals adapting to environments, and adapting to social interaction in groups and societies.

3. Technology Promoting Real-Life Social Interaction

Technologies can serve as media through which humans interact. The characteristics of these media affect the character of the interaction. For instance, using e-mail rather than face-to-face conversation frees the participants from many restrictions on time and space. The practice involves different modalities and has a narrower affective bandwidth (channel capacity for carrying emotional colouring of a message).[3] Emotional recognition, expression and synthesis are important areas where agent technology affords great opportunities to enhance our interactions with computers and with other humans through electronic media (Picard 1997). E-mail supports only relatively slow interactions, and thus the participants are given an opportunity to take time to think before responding, and may delay decision processes. This kind of

interaction creates a particular kind of social realm in space and time in which humans may interact.

3.1 *Agents extending the social self*

Using agents as surrogates or enabling technology for interaction among humans (or among humans and agents) may change the meaning of utterances such as: "I talk to him", "We've made this decision", and other expressions in which we refer to ourselves or our social groups.

3.2 *How will we change?*

How will we change when dealing with technological agents rather than human beings, e.g. software secretaries, avatars, beings for whom we do not know whether they are representatives of humans or virtual characters and talk-bots, or with humans to whom we are represented in new bodies or new modalities? Several issues to be aware of include:

Disadvantages

- Limited modes of interaction: e.g. concentration on written language as the primary source of interaction for socializing via computer. Similarly, interface design will determine the dominant modes of interaction with advanced social agents.
- Disconnection from society: frequent social interactions via computer networking within a world-wide social Web of friends, while separating or tending to ignore the immediate social 'real' environment. Complex synthetic social agents could act as pleasant or even addictive substitutes for interactions with humans, without danger of offending actual human interaction partners.
- Loss of human relatedness: relationships with remote friends can be close, but can be strongly selective, e.g. contact can be avoided easily in conflicts or difficult situations. Social agents can be built so that they feed the user's expectations and desires, which human agents might not be able to meet or satisfy.
- Social maladaptiveness: during interaction with agents and internet-working with people new norms and conventions are quickly formed. These new forms of interaction could misguide interaction in the real

world. Persons who are perceived as highly social in the virtual world, and/or in their interaction with social agents, might turn out to be socially handicapped in face-to-face interactions with humans.

- Body image changes: Due to the plasticity of our nervous system and whole body, sensori-motor coordination can be strongly influenced by interaction with agents, in the same way as the use of virtual environments has been shown to affect human sensori-motor coordination. Similar effects are expected in long-term interactions with social agents. For example: an important aspect in human-human communication is facial expressions, which serve as important social cues. The ability to express a range of emotions, and to recognize and distinguish subtle changes in facial expressions (and also other non-verbal cues involving body language) is a highly elaborated human social skill. In order to make social agents 'life-like' they can be given facial expressions which simulate animal or human ways to express emotions, but they are usually not achieving the subtlety and complexity of human facial expressions. Humans spending much time in interactions with social agents might therefore develop impoverished emotional perception and expression skills.

Advantages

- Extension of abilities: allowing new forms of working, cooperation, and socializing with people from different cultural and social backgrounds that are difficult to encounter in 'real life'. Social agents which can for instance help in intercultural communication (e.g. by having sufficient language skills and rudimentary knowledge of social interaction protocols in a particular culture) can support the formation of new communities.
- Social augmentation: agents and personal representatives can allow us to extend and multiply our single physical presence into many virtual entities. Also, different roles and personalities can be experienced in virtual, consequence-free environments. Interaction with social agents, if complemented with sufficient experience with social interactions in real life, can raise our awareness about the importance of personal social interactions and the appropriateness of different modes of communication and interaction.
- Reduced cognitive and interactive load: agents can reduce our cognitive load in particular for routine or stereotypical work which makes a huge part of normal working hours. Not all of our daily social interactions

require individual empathy and complex forms of social understanding which humans are experts in and which are difficult to model or simulate. But social agents can help in many routine and rather 'anonymous' interactions (e.g. handling requests for reprints, appointments, etc.).

The above lists are neither exhaustive nor meant as absolute judgements. Whether a factor is considered a disadvantage or advantage, and to what degree this is so, depends on each particular application and trade-offs in its particular cognitive, social, and cultural context.

3.3 *Agent authority*

When artificial agents act on our behalf, we are permitting them control over some of the authority that we might otherwise exercise. In a given situation, by letting an agent act on our behalf, it makes sense for the user to understand in what capacity the agent is allowed to act, and for a designer to be aware of the degree of authority the agent is given and of what impact this might have. The following list indicates several types of social agent roles with respect to authority and autonomy:

Levels of agent authority/autonomy:

- assistant (e.g. comparable to a 'secretary', performing well-defined standard or routine work, checking back with user in case of decision-making)
- associate (sharing workload, being able to solve relatively complex tasks on its own, can include ability to learn from the user, e.g. by demonstration)
- playmate, pet (autonomous agent, but having the particular role of entertaining and pleasing the user)
- guide (having a particularly strong expertise in one field and guiding the user in this domain, e.g. as a navigation aid in virtual environments)
- monitoring human performance (observing and monitoring human performance, ability to make suggestions, e.g. on efficiency improvements)
- human out of decision loop (independently acting without user feedback, possibly with the danger of acting 'out-of-control')

One dimension of concern here is where the actual integration of agent input

occurs. Does the agent interface with a human at a cognitive level, e.g. in terms of giving advice, or at a behavioural level, in which the agent represents or stands in for a human user or decides for the human which direction to go in (as might a driving agent in one's automobile)?

Designers should ask and users should want to know:

- What level of agent authority is generally appropriate for this application?
- Is the user aware of the level of authority she is giving the agent in each case?
- Should the user be made aware of agent actions?
- Is the task so trivial or the answer so obvious that it is not even worth asking the user?
- When is it appropriate (if ever) for the agent to interrupt the user with this particular information?
- Are the agent's actions at an appropriate level of authority in every specific case?
- What kinds of agent actions do or not require user approval?

The answers to such questions will depend on the candidate users and the particular purposes and tasks for which the agent is designed or employed. A thermostat agent might regulate the temperature in the user's office, keeping it within an acceptable range without ever or only rarely interacting with the user directly. Yet a user might wish to be consulted before a negotiation agent spends some money on a book or airline tickets for her, and almost certainly for a house, or before accepting a proposal of marriage or agreeing to an organ donation.

3.4 *Agent complexity*

There is an important relationship between human expectation and agent complexity: When presenting agents as interesting or as having exciting personality, a designer needs to consider the match between the user's conception of the agent's abilities and what these abilities actually comprise. If an agent turns out not to be as human-like as it might first appear to us as users, we often feel cheated or disappointed by what becomes an annoying complexity in appearance, e.g. 'fake cuteness' can be unpleasant and an anthropomorphic façade disappointing in the wake of 'unintelligent intelligence' or

irrelevant and useless advice. Users may be frustrated or feel resentment when a software agent fails to live up to expectations. Anthropomorphizing an agent can lead to this result if the designer carelessly leads the user to read too much in by representational choice. In contrast, if an exciting agent represents itself in a too dull or too simple manner, this may lead a user to ignore it. But a low complexity appearance that hides a more complex agent can be surprising, perhaps even pleasantly so, if the agent is found to be interesting or useful. If appearing too simple then it might fail to compete for the user's attention, resulting in unsuccessful interaction. For precise tasks with narrow domain or low autonomy appropriate, anthropomorphic representation may disturb a user, who may become uneasy with the perceived (or real) freedom allowed to an anthropomorphic agent.

Thus, one can identify a general correspondence principle for agent design (which is not likely to be without exception, especially when deception is desired by designers): Complexity of appearance in balanced design should reflect complexity of behaviour. As an agent, one should project one's abilities. A motto for agents then might be: *ADVERTIZE YOUR DEGREE OF COMPLEXITY!* This can be viewed as a particular corollary of a general principal in user interface design: in building an interface to an interactive system, a designer should preserve as much as possible the structure and semiotics of the source domain in the structure and semiotics of the user's target domain, where interaction is situated (Goguen 1999).

4. Uses of Technology

No one foresaw the pervasive influence that the telephone or Internet would have in our daily lives. As designers and users, we want to have a clear idea about what a particular agent is and what is it for. Building an agent resembles creating an artwork, which can have different meanings for different members of its audience. One cannot be sure what the created artifact will mean, or how it will be interpreted and received. What we design technology for is *not* necessarily what it will be used for.

Moreover, once a piece of social agent technology enters the world of the user, who is already embedded and situated in a physical and social environment, interaction may lead to uses beyond what anyone could have predicted. This is true even when there was a shared intention for designer and user in the

requirements for the technology. What is expressed in communication between those who design systems and those who use them may have little bearing when the agent technology is deployed and decisions are made at an emotional and embodied level.

5. Summary

Cognitive Technology is concerned with optimization of the 'cognitive fit' between humans and their tools. The advent of agent technology presents a vast realm in which agent designers have the opportunity to empower other human beings (and themselves) by appropriate design or to cause untold frustrations, inconvenience and misery to users who employ agents in daily life and social interaction if design should fail to be humane. This chapter raises a few issues of what is at stake in regard to integrating technology in a manner that helps people and respects human wholeness. Issues of longer term change where agent technology can shape interaction include social augmentation; changes in community, identity, place; processing speeds and modalities of interactions, and especially impacts on human cognition and social interaction. Awareness issues for appropriate requirements capture and design decisions, depending on the particular application of agent technology, include types and domains of authority permitted to agents, human expectation and agent complexity, social affordances, and cognitive and interactive load. Considerations for achieving the potential of agents in social technology and human-computer interactions, and the design of particular agents, should be informed by awareness of the issues that surround the problems of optimizing the social and cognitive fit between humans and their tools.

Technology already cognitively shapes and may in the very near future shape much more definitely the way in which humans represent and mirror conceptions of themselves. Many aspects of 'consciousness' like intentionality, agency, autobiographic (narrative) memory, monitoring and control of one's own behavior and the behavior of others are already modelled, operationalized and simulated, in one way or the other, in contemporary tools. This in turn directly affects the way we understand and operationalize what consciousness is for the mind, and it changes the way we assess the role and functions of consciousness.

Acknowledgements

Some of the ideas arose in the course of dynamic discussion on the issues presented here in the AAAI symposium workgroup 'Agent Technology and Human Cognition'. The other participants contributed significantly to the ideas we presented herein. We are grateful to them for their instigations and stimulating insights. They are: Dolores Cañamero, John Demiris, Simon Goss, Chisato Numaoka, Simon Penny, and Walter Van de Velde. The opinions expressed in this chapter are however our own.

Notes

1. This article summarizes and extends the contents of the discussion group 'Agent Technology and Human Cognition' organized by the authors which took place on 9th of November 1997 at the Massachusetts Institute of Technology as part of the American Association for Artificial Intelligence's Fall Symposium on Socially Intelligent Agents (SIA).

2. These issues mentioned in this paragraph are also the concerns of an international scientific organization, the Cognitive Technology Society (CTS).

3. The recent widespread use of simple ASCII graphics, so-called 'smileys' (collected by Sanderson 1993) is indicative of what issues are at stake. For example the character pairs :) and :(are symbols used to affectively mark e-mail or other text-based communications that provide some means to cope with the limitation on affective bandwidth in these media. Without such markings, the affective intentions of a writer can easily be misunderstood (as serious, for example, when joking, or as viciously critical when trying to be helpful). This may lead to what seem to the sender as inappropriate responses from the recipient. Such exchanges may escalate as can be frequently witnessed for example in newsgroup 'flame wars'. Such situations are frequently experienced by novice users (and less occasionally by experienced ones) of various text-based interpersonal communication media.

References

Cognitive Technology Society (CTS): http://edutools.cityu.edu.hk/ctsoc/

Dautenhahn, K. 1997. The Role of Interactive Conceptions of Intelligence and Life in Cognitive Technology. In J. P. Marsh, C. L. Nehaniv, and B. Gorayska (Eds.), *Proceedings of the Second International Conference on Cognitive Technology (August 25–28, 1997 – Aizu, Japan)*, IEEE Computer Society Press, 33–43.

Dautenhahn, K. 1998. The Art of Designing Socially Intelligent Agents: Science, Fiction and the Human in the Loop. *Applied Artificial Intelligence Journal, Special Issue on Socially Intelligent Agents.* 12 (7–8): 573–617.

Dautenhahn, K. 1999. Embodiment and Interaction in Socially Intelligent Life-Like Agents. In C. Nehaniv (Ed.), *Computation for Metaphors, Analogy and Agents,* (Lecture Notes in Artificial Intelligence, 1562), Springer Verlag, 102–142.

Goguen, J. 1999. An Introduction to Algebraic Semiotics, with Application to User Interface Design. In C. Nehaniv (Ed.), *Computation for Metaphors, Analogy and Agents,* (Lecture Notes in Artificial Intelligence, 1562), Springer Verlag, 242–291.

Gorayska, B. & J. Marsh 1996. Epistemic Technology and Relevance Analysis: Rethinking Cognitive Technology. In B. Gorayska and J. L. Mey (Eds.), *Cognitive Technology: In Search of a Humane Interface.* (Advances in Psychology, 113. Amsterdam: Elsevier/North Holland, 27–39.

Gorayska, B. & J. L. Mey 1996. Of Minds and Men. In B. Gorayska and J. L. Mey (Eds.), *Cognitive Technology: In Search of a Humane Interface.* (Advances in Psychology, 113). Amsterdam: Elsevier/North Holland, 1–24.

Gorayska, B., Marsh, J., & J. L. Mey 1997. Putting the Horse before the Cart: Formulating and Exploring Methods for Studying Cognitive Technology. In J. P. Marsh, C. L. Nehaniv, and B. Gorayska (Eds.), *Proceedings of the Second International Conference on Cognitive Technology (August 25–28, 1997 – Aizu, Japan)*, IEEE Computer Society Press, 2–9.

Haraway, D. 1991. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Centery. In D. Haraway, *Simians, Cyborgs and Women: The Reinvention of Nature*, New York: Routledge, 149–181.

Mey, J. L. 1997. Personal communication, presentation at the Second International Conference on Cognitive Technology, 25 August 1997 at the University of Aizu, Japan.

Nehaniv, C. L. 1997. Algebraic Models for Understanding: Coordinate Systems and Cognitive Empowerment. In J. P. Marsh, C. L. Nehaniv, and B. Gorayska (Eds.), *Proceedings of the Second International Conference on Cognitive Technology (August 25–28, 1997 – Aizu, Japan)*, IEEE Computer Society Press, 147–162.

Picard, R. 1997. *Affective Computing*. MIT Press.

Sanderson, D.W. 1993. *Smileys*, O'Reilly Publishing Company.

Name Index

A

Abelson, R.P., 62, 83, 223, 226, 257, 292, 299
Ackermann, E., 81
Agre, P.E., 13, 16, 23, 24, 197, 198, 221, 223, 225, 257, 399, 413
Alexander, K., 67, 81
Allen, J., 115, 134, 260
Allen, R., 320
Allie, P., 111
Anantharaman, T., 25
Anderson, A., 57
Anderson, J.R., 225, 257
Anderson, S.V., 257, 258
André, E., 103, 110
Andrews, T., 55
Anwar, A., 118, 133, 134
Aristotle, 156, 159, 161
Arkin, R.C., 202, 221
Armon-Jones, C., 147
Arnold, M.B., 236, 244, 258, 259
Asakawa, G., 34, 56
Ashby, W.R., 138, 161
Averill, J.R., 147, 161, 236, 259
Aylett, R.S., xx, 197ff

B

Baars, B., 114, 118, 134
Badler, N., 111
Baker, S., 1, 24
Balabanovic, M., 321
Balint, M., 298
Ballachey, E.L., 248, 259
Barden, J., 134
Bargh, J.A., 225, 258, 259
Barnes, D.P., xx, 197ff
Bartlett, F.C., 62, 81
Bates, J., 5, 8, 24, 25, 64, 81, 352, 396
Beaudoin, L., 164, 174, 176, 195
Bechara, A., 237, 257, 258
Bechdel, A., 347
Becker, H., 247, 258
Beckers, R., 213, 222, 405, 413
Beckwith, R., 82
Bederson, B., 92, 111
Belin, A. d'A., 383, 384, 393
Belitt, B., 58
Belk, R.W., 56
Bellman, K.L., xxii, 301ff
Benston, M.L., 37, 40, 56
Bergman, A., 299
Bers, M., see Umaschi Bers
Bettelheim, B., 63, 81
Bickmore, T., xviii, 7, 26, 85ff
Bier, E., 98, 103, 111
Bigelow, J., 227, 260
Billinghurst, M., 110
Blair, D., 4, 24
Bledsoe, W., 1, 24
Blumberg, B., 6, 10, 24, 97, 99, 102, 110, 351
Bly, S., xviii, 85ff
Bobrow, D.G., 321
Bogner, M., xviii, 113ff
Bonasso, R., 219, 222, 223
Bond, A.H., 309, 320
Boone, G., 115, 134
Booth, K., 57
Borovoy, R., 67, 81
Brahm, G.Jr., 57
Braitenberg, V., 305, 320
Bratman, M.E., 225, 227, 258
Bromley, D., 81
Brooks, K.M., 64, 81

Brooks, R.A., 1, 6, 18, 21, 24, 197, 198, 199, 202, 219, 222, 391, 393, 399, 413
Brown, J., 39, 56
Bruckman, A., 83, 327, 343, 347
Bruner, J., 2, 3, 4, 24, 62, 81
Brykov, M., 349
Brzezinski, J., 25
Buckley, W., 260
Burn, H., 82
Butterworth, G., 287, 298
Byrne, R.M.J., 225, 259

C

Caldwell, B., 109, 110
Caldwell, D.G., 221
Campbell, J., 63, 82
Campbell, L., 110
Campbell, M.S., 25
Cañamero, D., xix, 137ff, 425
Carbonell, J., 24
Cariani, P., 399, 413
Cassell, J., xviii, 28, 35, 38, 43, 55, 56, 57, 61ff, 99, 110
Castelfranchi, C., xxi, 225ff
Cesta, A., 245, 258
Chalmers, D.J., 171, 195
Chang, K., 110
Chapman, D., 23, 24, 221, 225, 257
Chater, N., 225, 260
Churchill, E., xviii, 85ff
Clark, M.S., 259
Clearwater, S.H., 245, 258
Cliff, D., 352, 356
Clore, G.L., 192, 195, 244, 260
Coddington, A.M., 206, 221, 222
Cohen, J., 385, 393
Cohen, P.R., 220, 222, 227, 240, 258, 259
Colby, K. M., 64, 82, 326, 347
Cole, M., 223
Collins, A., 192, 195, 244, 260
Collins, R., 156
Connell, J., 199, 219, 222
Conte, R., 227, 258
Cook, L., xviii, 85ff
Cooper, J., 247, 258
Coplien, J.O., 322
Cornelius, R.H., 144, 161
Courchesne, L., 396
Covi, L., 111
Craik, K., 184, 195
Cranach, M. von, 247, 25
Crary, J., 397, 413
Croucher, M., 174, 195
Crutchfield, R.S., 248, 259
Csikzentmihalyi, M., 55, 56
Curtis, P., 327, 347
Cutler, R., 109, 110

D

Daio, P., 222
Daiute, C., 64, 82
Dalla Volta, A., 247, 258
Damasio, A.R., 146, 161, 167, 176, 178, 179, 195, 236, 237, 257, 258
Damasio, H., 237, 257, 258
Dantzer, R., 146, 161
Dautenhahn, K., xv, xxiv, 4, 18, 25, 26, 140, 161, 162, 163, 301, 309, 310, 320, 415ff
Davidson, P., 258
Davis, P.K., 321
Davis, R., 207, 222
DeMaso, D., 81
Demazeau, Y., 300
Demers, L.P., 377, 378, 385, 388, 390, 392, 393
Demiris, J., 425
Deneubourg, J.-L., 222, 405, 413
Dennett, D., 11, 25, 176, 195
Dewey, J., 56, 248, 249, 258
D'Inverno, M., 225, 259
Dittmar, H., 35, 38, 56
Domeshek, E., 68, 82
Don, A., 3, 25, 63, 82
Donath, J.S., 31, 32, 33, 55, 56, 90, 110
Donegan, B., 81
Doyle, J., 225, 258
Dreyfus, H., 399, 413

Driscoll, M., 57
Drucker, J., 413
Druin, A., 30, 35, 45, 53, 56, 67, 82, 92, 111
Durkin, K., 58
Dweck, C.S., 225, 258, 259
Dyl, J., 30, 31, 45, 56, 67, 82

E
Eccles, R., 94, 111
Ekman, P., 162
Elliott, C., 3, 25
Ellsworth, P.C., 236, 260
English, A., 248, 259
English, H., 248, 259
Ephrati, E., 202, 222
Epstein, S., 225, 259
Erikson, E.H., 63, 82
Etzioni, O., 26
Evans, J.St.B.T., 225, 259
Evard, R., 327, 347
Everett, H., 285, 298
Eysenck, M.W., 110

F
Farhoodi, F., 220, 222
Farmer, D., 134, 383, 384, 393
Fellbaum, C., 82
Fikes, R., 260
Finke, L., 39, 56
Firby, R.J., 219, 222
Fishkin, K., 98, 103, 111
Fiske, S.T., 259
Flores, F., 65, 83
Foner, L.N., xxii, 44, 56, 65, 82, 323ff
Forsythe, C., 110
Frank, A., 9, 25, 26
Franklin, S., xviii, 113ff, 174, 193, 195, 381, 382, 393
Frenkel, K., 38, 56
Freud, A., 63, 82, 298
Frijda, N.H., 144, 161, 236, 244, 259
Fujita, M., 23, 25
Fuller, M., 63, 82
Fung, H., 83

G
Galanter, E., 227, 260
Gale, A., 110
Galotti, K.M., 225, 259
Galyean, T., 3, 25, 97, 102, 110
Garlan, D., 303, 320, 322
Garvey, C., 56
Gasser, L., 300, 309, 320
Gat, E., 201, 204, 222
Geldof, S., 162
Genesereth, M., 140, 161
Genovese, V., 199, 222
Gent, R. van, 351
Georgeff, M.P., 225, 227, 260
Gerth, H.H., 299
Ghanea-Hercock, R.A., 221, 222
Gilbert, D., 381, 393
Gilbert, J.H.V., 58
Gist, P.L., 225, 261
Glos, J., xvii, 27ff, 67, 82
Goffman, E., 92, 110
Goffman, I., 31, 57
Goguen, J., 423, 426
Goldberg, L., 310, 320
Goldstein, J., 34, 36, 57
Goldstein, K., 25
Goleman, D., 183, 195
Gollwitzer, P.M., 225, 226, 258, 259
Gombert, J.E., 74, 82
Gonzalez, J.A., 28, 31, 57
Gonzalez-Heydrich, J., 81
Gorayska, B., 416, 417, 425, 426
Goss, S., 425
Gould, J., 261
Graesser, A., 113, 114, 133, 134, 135, 381, 382, 393
Grand, S., xxii, 349ff
Grant, T.J., 220, 223
Gray, J.O., 221
Greenberg, M.L., 222
Greenberg, S., 92, 110
Gribble, S., 57
Grice, H.P., 276, 299
Grose, J., 110
Gross, D., 82

Grosz, B.J., 259
Gruber, T., 387, 393
Gugler, B., 247, 258

H
Habermas, J., 299
Hall, W., 81
Halpern, J.Y., 322
Hampson, S.E., 100, 110
Haraway, D., 415, 426
Harel, I., 82
Harnad, S., 198, 223
Harr, R., 161
Harro, V., 92, 111
Hart, D.M., 222
Hascombe, B., 40, 57
Haviland, J. M., 161
Hayes-Roth, B., 81, 200, 219, 223, 303, 317, 321, 351
Hewitt, C., 309, 321
Hickman, M., 82
Higgins, E.T., 225, 244, 259, 260, 261
Hilf, F.D., 347
Hillestad, R., 321
Hilton, J.L., 225, 261
Hobbs, J.R., 225, 259
Hodgson, P., 110
Hoffman, L.J., 384, 393
Hofstadter, D.R., 25, 115, 118, 134
Hogg, T., 245, 258
Holland, J.H., 119, 134
Holland, O.E., 222, 405, 413
Hollander, E.P., 247, 259
Holyoak, K., 134
Homans, G.C., 247, 260
Hoogstra, L., 83
Horrobin, A.J., 221
Houde, S., 30, 57
Howe, R.E., 222
Hsu, D., 57
Hsu, F., 1, 25
Huberman, B.A., 245, 258
Huberman, M., 323, 348
Huhns, M., 300
Husbands, P., 376
Hutchens, J., 393

I
Indermuhle, K., 247, 258
Inkpen, K., 40, 57
Inman, J., 309, 321
Ishii, H., 29, 55, 57, 58, 67, 82
Israel, D.J., 227, 258

J
Jackson, J., 115, 119, 124, 134
Jacobson, E., 299
Jacquet-Droz, 385
James, E., 13, 25
Jenkins, H., 40, 55, 57, 63, 82
Jennison, L., 222
Jessell, T.M., 161
Johnson, C.E., 58
Johnson, M., 399, 413
Johnson, P., 322
Johnson, W.L., 57, 161, 348
Johnson-Laird, P.N., 195, 225, 261
John-Steiner, V., 223
Johnston, O., 8, 26
Johnston, W.L., 26
Jones, E., 237, 259
Jones, G., 82
Jones, S.D., 219, 221
Jung, C.G., 286, 299

K
Kac, E., 379, 393
Kageyama, K., 23, 25
Kahneman, D., 225, 261
Kalbermatten, V., 247, 258
Kalin, N.H., 147, 161
Kamptner, N.L., 57
Kandel, E.R., 145, 161
Kanerva, P., 115, 134
Karlin, J., 81
Karmiloff-Smith, A., 74, 82
Kauffman, S., 405, 413
Keane, W., 138, 159, 161
Kellerman H., 161
Kellner, D., 31, 57

Kemper, T.D., 156, 161
Kennedy, J., 302, 321
Kernberg, O., 278, 299
Ketchpel, S., 140
Kiczales, G., 315, 321
Kincaid, J., 57
Kingery, W.D., 56
Klawe, M., 57
Kluckhohn, C., 248, 259
Koehler, W., 247, 259
Kohut, H., 278, 284, 299
Kolb, W.L., 261
Kollock, P., 56, 110
Kolodner, J., 115, 134
Kortenkamp, D., 223
Kozierok, R., 115, 134, 333
Kramarae, C., 37, 56, 57, 58
Krech, D., 248, 259
Kube, C., 199, 223
Kunda, Z., 225, 259
Kuniyoshi, Y., 223

L

Lacan, J., 299
Lalanda, P., 321
Landauer, C., xxii, 301ff
Langheinrich, M., 26
Lashkari, Y., 115, 134
Lathers, D., 101, 111
Laurel, B., 3, 25, 63, 82
Lawry, J., 57
Layzell, P., 222
Lazarus, R.S., 236, 259
LeDoux, J.E., 145, 161
LeDoux, J.E., 168, 176, 183, 195
Leggett, E.L., 225, 259
Lenting, J.H.J., 220, 223
Leonard, A., 382, 393
Leroux, S., 57
Levesque, H.J., 227, 240, 258
Levy, S., 385, 393
Lewis, M.G., 37, 57, 161
Lifton, R.J., 31, 58
Linde, C., 63, 82
Lindzey, G., 260
Lipman-Blumen, J., 38, 58
Little, T.D.C., 81
Lock, A., 298
Lombardi, J.A., 321
Loyall, A.B., 352
Loyall, B., 7, 8, 9, 24, 25
Lualdi, S., 81
Lubar, S., 56
Luck, M., 225, 259
Luger, G.F., 312, 322
Lutz, C.A., 147, 161

M

MacIntyre, A., 252, 259
Maes, P., 6, 8, 25, 81, 98, 99, 111, 222, 304, 315, 322, 413
Magni, M., 222
Magritte, R., 403
Mahler, M., 283, 284, 299
Maillardet, 385
Malhotra, A., 356
Mandler, G., 238, 259
Mani, K., 55
Marsh, J., 416, 417, 425, 426
Marsland, T.A., 25
Martin, A., 81, 83
Martlew, M., 36, 44, 58
Maslow, A.H., 247, 260
Massumi, B., 16, 25
Mataric, M., 161
Matlack, C., 1, 24
Maturana, H.R., 114, 134, 149, 161
Mauldin, M., 64, 82, 326, 327, 328, 336, 339, 341, 348
Maxwell, J.A., 323, 348
McCabe, A., 36, 48, 58
McCauley, L., 116, 117, 133, 134
McDermott, D., 225, 260
McGaugh, J., 248, 258
McLuhan, E., 58
McLuhan, M., 58, 379, 393
McPherson, T., 57
Mead, G.H., 286, 299
Medin, D.L., 225, 260
Metral, M., 134

Mey, J.L., 415, 416, 417, 426
Meyer, J.-A., 25, 161, 376
Meyer, T., 4, 24
Meystel, A., 322
Miceli, M., xxi, 225ff
Miedaner, T., 12, 25
Miles, M.B., 323, 348
Miller, P., 62, 63, 83
Miller, A., 299
Miller, G.A., 68, 82, 227, 260
Miller, K., 68, 82
Miller, L.H., 316, 322
Miller, W., 111
Milne, A.A., 45, 47, 58
Minsky, M.L., 6, 26, 176, 195, 199, 223
Mintz, J., 83
Mitchell, M., 115, 118, 120, 134, 135
Mitchell, W.J.T., 83
Mithen, S., 164, 176, 195
Montford, N., 64, 83
Moon, Y., 94, 111
Moore, R.C., 225, 259
Morgan, J., 259
Morgenstern, O., 299
Morignot, P., 321
Müller, J., 103, 110
Müller, J.-P., 300
Murphy, G.L., 225, 260
Murphy, R., 223
Murray, J., 64, 83

N

Nardi, D., 315, 322
Ndunda, M., 57
Neal Reilly, S., 6, 26
Negatu, A., 117, 133, 135
Nehaniv, C.L., xxiv, 415ff
Neruda, P., 58
Neumark, N., 393
Newell, A., 139, 161
Newman, J.E., 322
Newstead, S.E., 225, 259
Nicholson, J., 55
Nilsson, N.J., 181, 195
Nisbett, R., 237, 259
Nohria, N., 94, 111
Noma, T., 111
Norman, D., 58
North, S., 326, 348
Nowatzyk, A., 25
Numaoka, C., 158, 162, 425

O

Oaksford, M., 225, 260
Oatley, K., 176, 195
Odetti, L., 222
Ohare, G., 300
O'Hare, J.J., 221
Olde, B., 135
Olson, J., 96, 111
O'Neill Brown, P., 15, 26
Opton, E.M.Jr., 236, 259
Ortony, A., 192, 195, 244, 260
Osman, A.A., 221

P

Papert, S., 45, 50, 52, 58, 65, 74, 81, 82, 83
Parsons, T., 247, 260
Pears, D., 238, 239, 260
Penberthy, J.C., 206, 223
Penny, S., xxiii, 389, 393, 395ff, 425
Pepper, S.C., 247, 260
Perry, R.B., 247, 260
Peterson, C., 36, 48, 58
Peterson, D., 194
Petta, P., 24
Pfleger, K., 321
Picard, R.W., 160, 162, 178, 195, 418, 426
Pine, F., 299
Plutchik, R., 161
Polanyi, L., 62, 83
Polkinghorne, D., 62, 63, 83
Pollack, M.E., 227, 258, 259
Polshaw, M.M.T., 221
Poor, R., 40, 55, 58
Porter, T., 10, 26
Poster, M., 33, 58
Potts, R., 83

Prentice-Dunn, S., 99, 111
Prevost, S., xviii, 85ff
Pribram, K.H., 145, 162, 227, 260
Pylyshyn, Z.W., 225, 260

Q
Quilici, A., 316, 322

R
Ramamurthy, U., xviii, 113ff
Rao, A.S., 225, 227, 250
Rao, R., 162, 223
Ratner, J., 110
Raybourn, E.M., 92, 111
Reilly, W.S., 6, 24, 26
Reinefeld, A., 281, 300
Reinhardt, C.A., 316, 320
Resner, B., 25
Resnick, M., 58
Resnick, R., 65, 74, 81, 83
Richards, N., 222
Riecken, H.W., 247, 260
Riesbeck, C., 65, 83
Rist, T., 110
Rocco, E., 111
Rochberg-Halton, E., 31, 56
Rogers, R., 111
Roitblat, H. L., 161
Rokeach, M., 247, 248, 260
Rosch E., 399, 413
Rose, J., 58
Rose, S., 167, 195
Rosen, S., 63, 83
Rosenblueth, A., 227, 260
Rosenschein, J.S., 202, 222, 223
Ross, M., 225, 260
Rostow Kuznets, L., 58
Rucker, L., 34, 56
Ryle, G., 172, 195

S
Sacks, O., 309, 322
Sakai, M., 36, 43, 44, 48, 55, 58
Salomon, G., 57
Salvatoriello, R., 25
Sanderson, D.W., 425, 426
Sandewall, E., 260
Sawhney, N., 143, 162
Schaeffer, J., 25
Schank, R., 62, 65, 83
Schank, R.C., 223, 292, 299
Schempf, H., 1, 26
Scherer, L. R., 162
Schiller, C., 322
Schmandt, C., .143, 162
Schmidt, D.C., 322
Schrooten, R., 162
Schulte, J., 404, 413
Schwartz, N., 244, 260
Schwartz, T., 161
Scribner, S., 223
Searle, H.F., 299
Sedighian, K., 57
Sengers, P., xvii, 1ff
Shakes, J., 1, 26
Shannon, C.E., 299
Shatluck, J., 57
Shaw, M., 303, 322
Sheridan, A., 299
Sheth, S., 25
Shiano, D.J., 91
Shrager, J., 326, 348
Sidner, C.L., 259
Sills, D.L., 258
Simon, H.A., 139, 161, 176, 195, 225, 260
Sims, K., 352
Singh, M. P., 162
Slagter, S., 92, 111
Sloman, A., xix, 163ff, 219, 223, 225, 260
Smith, B.C., 315, 322
Smith, C.A., 236, 260
Smith, M., 56, 110
Smith, R.G., 222
Solomon, C., 56
Song, H., 116, 132, 135
Sontag, S., 24, 26
Sorrentino, R.M., 225, 259, 260, 261
Souberman, E., 223

Spitz, R., 299
Spurr, K., 222
Steels, L., 7, 26, 223, 226, 261
Stern, A., 25
Stern, C., 322
Stewart, J., 92, 111
Stewart, S., 58
Stone, B.A., 7, 26
Stone, M., 98, 103, 111
Strohecker, C., 81
Strommen, E., 58
Strommen, E., 67, 81
Stubbs, B., 384, 393
Suchman, L., 65, 83, 197, 198, 223
Sullivan, J.W., xviii, 85ff
Sutton-Smith, B., 34, 35, 37, 58
Sycara, K.P., 25, 116, 135

T
Taha, L., 109, 110
Thomas, F., 8, 25, 26
Thompson, E., 399, 413
Thorisson, K., 100, 103, 111
Tomkins, S.S., 145, 146
Tranel, D., 237, 258
Trappl, R., 24
Tsukune, H., 223
Turing, A.M., 324, 325, 348
Turkle, S., 32, 37, 58, 65, 81, 83, 91, 111
Turner, V., 62, 83
Tuteja, M., 55
Tverski, A., 225, 261

U
Uexkull, von J., 308, 322
Ullmer, B., 29, 55, 57, 58, 67, 82
Umaschi Bers, M., xviii, 35, 42, 53, 57, 59, 61ff
Upitis, R., 57

V
Van de Velde, W., xix, 137ff, 425
Vander Zanden, J., 95, 111
Varela, F., 114, 134, 149, 161, 399, 413
Vaucanson, 385
Vere, S., 7, 26, 111
Vertegaal, R., 92, 111
Vilhjalmsson, H., 110
Vogt, P., 223
Von Neumann, J., 299
Vorn, B., xxiii, 377ff
Vygotsky, L.S., 198, 223, 276, 299

W
Walter, D.O., 304, 310, 320
Wan, Y., 135
Wapner, S., 30, 31, 45, 56, 67, 82
Wason, P.C., 225, 261
Weber, M.M., 296, 297, 299
Weber, S., 347
Weizenbaum, J., 64, 83, 165, 195, 325, 326, 348
Weld, D., 206, 223
Werner, E., xxi, 263ff
Wheeler, M., 113, 135
White, S., 91, 98, 111
Wiener, N., 167, 195, 227, 260
Wiggins, B.B., 95, 111
Wiggins, G.A., 95, 111
Wigren, J., 63, 83
Williams, M., 53, 59
Williams, R.M.Jr., 247, 248, 261
Wilson, I., 55, 57
Wilson, S.W., 25, 376
Winnicott, D.W., 30, 37, 59, 67, 83, 277, 280, 300
Winograd, T., 65, 83
Wise, J.M., 16, 26
Wittenbrink, B., 225, 261
Wittgenstein, L., 169, 195, 300
Wolcott, H., 323, 348
Wooldrige, M. J., 25, 162, 223
Wresch, W., 82
Wright, I., 174, 186, 195

Y
Yan, H., 110
Yeo, R., 322

Z
Zelinsky, A., 204, 223
Zellweger, P., 103, 111
Zeng, D., 135
Zhang, H., 199, 223
Zhang, Z., 117, 119, 121, 133, 135
Zingrone, F., 58

Subject Index

A

abandonment, 273
action, 227
action-expression, 26
action-selection, 6, 23, 25, 114, 115, 118, 119
active integration processes, 315
active social space, 295
actor, 101
adaptation and emotion, 145, 150, 154
adaptation, 138, 140, 145, 310
adaptivity of behavior, 304
aestetic of behavior, 398
affective bandwidth, 418, 425
affective, 160
agency, 402
agent software, 142, 144, 155
agent and agency, 140, 148
agent architectures, 199
agent authority, 422
agent complexity, 422, 424
agent design, 395
agents, 302, 377, 379, 381, 382, 383, 389, 390, 391
AI Lab Olympics, 347
AI planning system, 201, 206, 219
alarm system, 176, 178, 180, 181
alternative AI, 6, 12, 15, 19
altruistic, 202
amygdala, 176, 183
analog vs digital, 181
analogy-making, 120, 129, 134
animal brains, 168, 169
animal intelligence, 163, 165
animal minds, 173, 175
animals, 178
animation, 8
anthropomorphic, 28, 34
anthropomorphism, 345, 346, 350, 386, 388
appearance, 138, 148, 149, 150, 154, 155, 157, 158
apprehension, 178
architectual requirements, 163
architectural layers, 175, 176, 178
architecture of vision, 177
architecture, 180
architecture-based concepts, 169
architectures explain, 175
architectures for love, 174
architectures, explanatory power, 175
architectures, varieties of, 175, 191
arena, 124, 125
arguing at cross-purposes, 180
Artificial Intelligence (AI), 64, 135, 163-165, 396
Artificial Intelligence, doubters and fearers about, 164
Artificial Intelligence, prejudice against, 164, 165
artificial life agents, 114
Artificial life (AL, Alife), 138, 398
artificial social agents, 298
artistic practice, 395, 397
artworks, 378, 385, 395
assistant, digital, 142
assistant, personal, 143
associative memory, 115, 118, 119, 184
atomization, 7, 19, 20, 21
attention, 113, 114, 118, 119, 124, 125, 352, 365
attention, control of, 174, 179, 180, 186
attitudes, 174
authorable, 89, 97, 102

authoring, 86, 87, 102, 107, 103
autonomous agent, 113, 114, 132, 134, 135
autonomous personal representative, 87, 95, 97, 99, 102, 108, 109
autonomous synthetic characters, 85, 86, 92
autonomy, 138, 140, 269, 279, 285, 290, 302, 386, 387, 388
autonomy and emotion, 145
autonomy and motivation, 145, 153
auto-pedagogic, 406, 409
autotopography, 31, 35
avatars, 86, 88, 91, 97, 99, 103, 104, 105, 106, 107
awareness and context perception, 143
awareness, 140, 142, 144, 160

B

barrier, 308
behavior network, 115, 118, 119, 132
behavior stream, 116, 129
behaviour (activity) selection, 138, 141, 145
behaviour packet, 204, 206, 208, 209, 210, 213, 216, 217, 220
behaviour pattern, 203, 206, 209, 211, 212, 214, 215, 216, 217, 219, 220
behaviour script, 204, 206, 208-211, 213, 215, 216, 217, 219, 220
behaviour selection and emotion, 146, 153
behaviour selection and motivation, 145, 153
Behaviour Synthesis Architecture (BSA), 202-206, 211, 219
behaviour, internal, 180
behavioural agent, 199, 202, 206, 207, 208
behavioural approach, 198, 199, 200, 208
behavioural architecture, 202, 204, 211, 219
beliefs, 227, 266, 269
believable characters, 64
believable, 396
biology, 192
blackboard, 127
body behaviors, 62
'bots', 379, 381, 382, 383, 391
brain abnormality, 191
brain damage, 163, 164, 175
brain science, 192
brain stem, 183
brains and machines, 165
brains, chemical controls, 176
broadcast manager, 124, 127
broadcast, 114, 118, 124, 125, 127, 129, 133
building blocks, 350, 354

C

Calendar Agent, 115
capture heads, 212, 214
case-based memory, 115, 120, 123, 124, 129, 130, 131
case-based reasoning, 115, 119, 122, 130, 131
causation in physical and virtual machines, 171
causation in virtual machines, 172
cells, 350
character individuation, 89, 103
chemicals and brains, 188
chemicals, 354, 357
cherished possessions, 30
childhood, 27, 28, 35, 36, 55, 63, 76, 83
classical AI, 6, 16, 19
classifier system, 119, 134
close-coupled cooperation, 211
cluster concepts, 169
coalition manager, 124, 126, 127
coalition, 114, 124, 125, 126, 127
codelets, 118, 119, 121, 122, 123, 124, 125, 126, 127, 129, 130, 131
cognitive agent(s), 114, 227
cognitive and interactive load, 420, 424
cognitive fit, 415, 417, 424
cognitive needs, 54
cognitive prosthesis, 118

cognitive prosthetics, 416
cognitive software agents, 114
Cognitive Technology (CT), 415, 416, 424, 425, 426
Cognitive Technology Society (CTS), 425
co-habited mixed reality, 138, 140, 141, 142, 143, 154, 160
cold mental processes, 179
Colin, 330, 337, 338
collision avoidance, 202, 212, 213
collocation, 92, 94
colonies of animals, 182
communication, 62, 64, 66, 67, 73, 77, 78, 79, 80, 83, 264, 275, 276, 277, 292
communicative behaviors, 67
community versus group, 141
community, 140, 143
communityware, 141
competence, 149, 154, 158
competition for attention, 141, 144, 148, 149, 158
complete agents, 164
complex systems, 301
composition workspace, 119
computational agents, 114
computational construction kit, 65, 83
computational models, 188, 192
computationally reflective, 315
computationally-augmented soft toys, 67
computer human interaction, 89
computer mediated collaborative work, 89
computer science, 192
computer viruses, 379, 381, 383, 384, 391
computers, 172
COMRIS agent model, 148
COMRIS parrot, 143, 160
COMRIS and information push, 144
COMRIS Personal Assistant (PA), 143
COMRIS Personal Representative Assistant (PRA), 143
COMRIS, 139, 141, 143, 154, 155, 158, 160
concept formation, 114
conceptual analysis, 163, 165
conceptual contexts, 114
conceptual learning, 113, 130, 131, 133
conceptual model of mind, 132
conceptually confused questions, 169
conflict resolution, 182
Conscious Mattie (CMattie), 113ff
consciousness, 113, 114, 118, 119, 123, 124, 125, 127-130, 133, 134, 424
constructed complex systems, 302
constructionism, 65
content addressable memory, 118
context and meta-management, 189
context, and reactions, 183
control strategy, 266
control systems, 113
control, 79, 80
control, chemical, 188
control, in minds and brains, 176
control, losing presupposes having, 174
control, loss of, 179
control, semantic and non-semantic, 188
controlled sources of variation, 313
conventions, 418, 419
conversational coherence, 94
conversational narratives, 66
conversational stories of personal experience, 62
conversational structure, 62, 69, 70, 76, 77, 78, 79
Convulsive Machine, 389, 390
cooperating robots, 214, 219
cooperation, 272, 273, 275-278, 281, 283, 284, 289, 290, 296, 297
cooperative act, 273, 296
cooperative linguistic behavior, 276
cooperative object exchange, 215
cooperative relocation, 210, 214, 215
copy internalization, 266
Copycat, 115, 120
coupling, structural, 138, 153, 154
coupling, sympathetic, 158
creatures, 350
Creatures, 350

crime and causation, 171
critical technical practice, 16
CSCW, 327
cultural contexts, 114
cultural values, 382, 384, 386
culture and meta-management, 190
culture, 4, 5, 15, 16, 18, 20, 21, 26, 177
cyberbiography, 87, 88
cyborgs, 415

D

Darwinism, 384
decision making, 114
definitions, 169, 180
deictic representation, 13
deliberation, 172, 177, 184
demons, 118, 124, 125
dependency, 273, 284, 285, 290
design decisions, 424
designs, 164
development, forms of, 190
dialogue space, 43
digital augmentation, 27, 54
disease, 353, 370
document avatars, 86, 88, 97
domain knowledge, 122
domain, 323, 324, 327, 345, 346
dormant dispositions, 174
dreaming, 369, 370
drives, 116, 118, 119, 181, 352, 366
dyad, 279, 289, 295
dynamic virtual architectures, 377

E

earthworms, 167
ecological niche, 304, 305
effective autonomy, 303
egocentrism, 78
egotistic, 202
ELIZA, 324, 325, 326, 327, 339, 347, 348
e-mail (email), 85, 88, 90, 93, 94, 95, 115, 116, 118, 120, 121, 122, 123, 124, 134, 415, 418, 425
embodied computer characters, 95
embodiment, 139, 304, 307, 308, 377, 398, 408
embodiment, computational, 301
emergence, 189, 350, 375, 405
emergency events, 120
emotion, definition of, 180
emotion, evolution of, 169
emotional affordances, 67
emotional and psychological value, 27
emotional appraisal, 236
emotional maturity, 191
emotional value, 38, 54
emotions (affective states, emotional states), 138, 144, 150, 153, 154
emotions and control, 174
emotions in infants and insects, 184
emotions in robots, 178
emotions, primary, secondary, tertiary, 178-180, 183, 189
emotions, 114, 116, 118, 119
emotions, and adaptation, 145, 150, 154
emotions, and communication, 146, 158
emotions, and homeostasis, 145, 153
emotions, and hormones, 151, 153
emotions, and motivation, 145, 146, 153, 155
emotions, and physiology, 138, 144, 149, 155
emotions, and power, 156, 157
emotions, and rhetoric 155, 156
emotions, and self-/other image, 147, 156, 157
emotions, and social interaction, 146, 154, 155, 156, 158
emotions, and social values, 147
emotions, and status 156, 157
emotions, basic emotions, 150, 154
emotions, bodily-grounded, 138, 154, 160
emotions, due to 'stupid filters', 186
emotions, emotional expression, 146, 154, 155, 157, 158
emotions, emotional memory, 158
emotions, functional view of, 144
emotions, in robots, 174

emotions, primitive, 168
emotions, three types, 178
emotions, peripheral, central, 179
empathy, 421
empowerment, 79, 80
enabling qualities, 234
energy, 166
episodic memory, 115, 122
Espace Vectoriel, 377, 378
evaluation(s), 180, 270, 272, 289, 296, 297
evaluation, and action, 239-244
evaluation, and emotion, 235-239, 244
evaluation, and rationality, 238-239
evaluation, and social action, 244-247
evaluation, and social comparison, 246
evaluation, and social image, 246-247
evaluation, and social power, 245-246
evaluation, and subjectivity, 237-239
evaluation, by classification, 233-234
evaluation, by performance, 233
evaluation, by standard, 234-235
evaluation, definition of, 230
evaluation, explicit, 240
evaluation, function of, 241-244
evaluation, hypothetical, 231, 243
evaluation, implicit, 240
evaluation, in preferences and choices, 243-244
evaluation, in problem solving, 242-243
evaluation, negative, 232
evaluation, of harmfulness, 232
evaluation, of insufficiency, 232
evaluation, positive, 232
evaluation, potential, 240
evaluation, processes of, 233-235
evaluative information, 266, 268
evaluative state, 265
evocative object, 32
evocative objects, 65
evolution of deliberation, 191
evolution of planning, 177
evolution, 163, 164, 168, 183, 353
evolution, copying in, 177
extended self, 282
externalization, 266, 270-274, 279, 280, 281

F

face-to-face communication, 85, 89, 92, 93, 99
facial expression, 100, 101, 102, 104
false self, 280
family resemblance concepts, 169
feeling, 165
feminist pedagogy, 39
feminist software design, 38, 39
flexibility, 183, 186
FORTRAN, 347
Fred and Ginger, 210, 211, 215, 216
free will, 174
Fugitive, 406ff
fun, 323ff
functions and evolution, 189

G

game playing machines, 172
gap between AI and animals, 165
gender-neutral, 28, 34, 35, 38, 40
'general purpose fish', 305
generalisation, 364
generative processes, 305, 310, 313
generator codelet, 125, 127
genes, 354, 359
gesture, 89, 265, 266, 267, 276, 294
gesture-action strategy, 267
ghost in the machine, 172
global alarms, and meta-management, 187
global alarms, and attention filter, 186
global alarms, and emotions, 189
global alarms, 183
global alarms, and deliberative mechanisms, 185
global workspace theory, 114, 115, 116, 118, 127, 128, 133
GNU Emacs Lisp, 326
goal adoption, 245
goal contexts, 114, 116
goal state, 201

goal, 227, 281, 292
goldfish cannot long, 173
good-enough adult, 277
graphical user interface (GUI), 29, 67, 69, 70
grief, 174
GroupWare, 92, 93, 94

H
Hamsterdam, 10, 24
heterogeneous multi-agent model, 197
historicity, 140
human cognition, 415, 416, 424
human reflection, 221
human wholeness, 417, 424
Human-Computer Interaction (HCI), 62, 64, 67, 416
humanities, 14, 21, 22
hybrid architecture, 181, 184

I
'I', 270
icons, 29
identity, 31, 32, 33, 34, 36, 37, 55, 61-66, 75, 77, 79, 83, 89ff, 111, 418, 424
identity, imagery, 409
imaginary friend, 37
immersion, 408
implementation and supervenience, 167
individual differences, 100, 102
infants, emotions in, 184
information processing architectures, 167, 193
information processing theories, 164
information processing, 163, 165, 168, 172, 179
information processing, science of, 170
information state, 265, 267, 293, 294, 295, 296
information, 166
informational content, 379, 382, 383, 384, 385, 386
infra-red video, 406
infrastructure, 303
infrastructure, reflective, 315
insects, 180
integration concept, 307
intelligence, 319
Intelligent User Support (IUS), 316
intention, 269, 298
intentional states, 263, 264, 269, 280, 296, 297
Interactive Space, 318
interactive story system, 36
interactive storytellers, 61, 66, 69, 79
intercultural communication, 420
interest models, 138
interest, 138, 144, 155, 156, 157
interest-based navigation, 141, 144, 148
interface, 409, 412
internal conditions, 116, 119
internalization operator, 268, 269
internalization, 263, 264, 266, 268-273, 275, 276, 279, 280, 281, 286, 287, 288, 290, 298
interpreting operator, 267
introspection as evidence, 191

J
Julia, 324, 327ff

K
keepsake object, 27, 28, 29, 30, 31, 34, 35, 38, 42, 47, 52, 54, 55
knowledge-basd polymorphism, 315
Kumo, 89

L
La Cour des Miracles, 377, 388
language, 353, 372
layered architectures, 175
layered sensory and motor systems, 177, 181, 182, 188
layers, 164
layers, reactive, deliberative, meta-management, 176
learning experience, 79
learning, 113-115, 119, 120, 122, 128-130, 131, 133, 134, 135, 164, 177, 184, 361

learning, forms of, 190
levels of description, 170
local decision-making, 200
local minimum, 201
Loebner Prize Competition, 348
longing, 173
loose-coupled cooperation, 215, 217
losing control of attention, 174
love and culture, 190
love, varieties of, 174
love, 163, 165, 168, 174, 192

M

Machine Culture, 396
machine vision, 406
machines, kinds of, 166
machinic behaviors, 391
MACTA (Multiple Automata for Complex Task Achievement), 200, 206, 208, 210, 217-220
mapping (a MUD), 335
market, 142
market, electronic, 142
master-slave hierarchy, 207
matter, 166
meaning, 62, 65, 71, 75, 79, 81, 397
meaningful containers, 30
mechanism, 165
media of communication, 379
meditation, 190
MegaHal, 393
memories, memory, 27, 28, 32, 33, 34, 36, 54, 114, 115, 118-125, 127, 129, 130, 131, 132, 134
memory, and serial access, 184
memory, associative, 176
mental concepts, 165
mere machines, 170
metacognition, 118, 119, 124
meta-interaction, 278, 279
meta-knowledge, 315
meta-management and context, 189
meta-management, advantages of, 186
micro-society, 200
minds and brains, 171,
minds, varieties of, 173, 175
Mission Organiser, 206, 209
MIT AI Lab, 325, 326, 347
modular architecture, 115
modularity and evolution, 191
Mood Analysis Engine, 408
mood, 188
MOOs, 86, 90, 91
mother-child relationship, 265, 290
motivation, 352, 367
motivations and behaviour selection, 152, 153
motivations and drives, 138, 152
motivations and homeostasis, 145, 151, 154
motivations, 138, 145, 150, 152, 153, 154
move-internalization, 273
MUD robot, 327
multi-agent system, 118
multimedia, 33, 34, 377, 392
multi-modal communication, 86, 97
multiple personality disorder, 189
multiple senses, 114
multi-robot actions, 210
multi-user domain, multi-user dungeon (MUD), 86, 90, 91, 306, 324, 327, 328, 329, 332, 335, 340, 343, 347
multi-user virtual environment (MUVE), 318
MUSH (multi-user shared hallucination), 331

N

Narrative Intelligence, 1, 3, 4, 5, 7, 9, 11, 13, 18, 19
narrative psychology, 3, 21
narrative voice, 63, 75
narrative, 2, 3, 4, 5, 7, 9, 10, 11, 13, 14, 17, 18, 19, 20, 21, 22, 23, 24
narratives, 27, 28, 34
natural language, 115, 116, 121, 129, 130, 131
negotiation, 209
neurones, 354, 355

non-functional mental states, 189
nonverbal behavior, 96, 101
norms, 263, 288, 290, 419
Norns, 350
nothing buttery fallacy, 170
novelists, 168

O

objectivity, 13
observer, 101
omnipotence, 282, 284
ontogeny, 263, 276, 286
ontological neurosis, 166
ontology, 171
organisms as information processors, 167, 170

P

pain, 186
Pandemonium theory, 124
parental mirroring, 276
PARRY, 324, 326, 327, 339
particularity, 9
partnership, 273
pass deflection, 340
perception module, 118, 119, 120, 121, 129, 130
perception registers, 116, 119, 122, 123, 129, 132
perception, 114, 116, 118-124, 129, 130, 132, 135, 172
perceptual contexts, 114
perceptual learning, 190
performance, 149, 150
periodic table of chemical elements, 175
personal histories, 31
personal meaning, 27
personal representation, 85, 89, 90, 91, 92, 93
personal representative, 85, 87, 88, 90, 95, 96, 97, 98, 99, 100, 102, 108, 109
personal stories, 61, 63, 65
personality, 86, 89, 100, 101, 110, 350, 352, 418, 422
perturbances, 179, 180
perturbances, emotional and unemotional, 180
Petit Mal, 399ff
phicons, 29
philosophy, 164, 171, 180, 192
physics, 170, 171
physiological parameters, 138, 151, 154, 158
physiology and secondary emotions, 178
physiology, synthetic 138, 150, 151
planning, 184
plans, evolved, 184
plans, in reactive systems, 184
plans, saved for re-use, 186
plants, 180, 182
play, 38, 40, 42, 44, 46, 47, 50, 52, 53
playing field, 124, 125, 126
PMS, 333, 340, 342
poetics of interaction, 410
poets, 168
Poplog, 192
posed problem, 316
poverty and causation, 171
power, 279, 281, 284, 285, 292
presentation, 138, 148, 150, 154, 155, 157, 158
primary emotions, 183
primitive actions, 207, 209
Problem Manager (PM), 316
problem-solving, 114
protean self, 31
psychodynamic interactions, 298
psychodynamic process, 270
psychology, 164, 180, 192

Q

qualitative analysis, 323
questionnaires, 379, 380, 391, 392

R

RANDOM-PARRY, 347
reactions, chained, 182
reactions, direct and indirect, 182
reactions, goal-using, 183
reactions, learnt vs innate, 183

reactions, selection of, 182
Reactive Action Packets (RAPs), 219
reactive architecture, 180
reactive systems, limits of, 181
reactive systems, varieties of, 181
reference internalization, 266
referentiality, 21
Reflective Agent (RA), 202, 206-209, 217
reflective approach, 197-201
refusal, 269, 283, 284
regression, 273, 274, 278, 279, 280
reinforcement, 362
relevance, 148, 149, 154, 158
relief after apprehension, 178
representation, 178, 264-269, 271, 274, 276, 277, 278, 286, 287, 293, 294, 377, 383, 388, 390, 392
reproduction, 371
robotic artwork, 399
robotic machines, 377
robotics, 377, 379, 385, 392
robots and meta-management, 193
robots and qualia, 193
robots and therapy, 193
robots, 165, 193, 378, 379, 381, 382, 385, 388, 391, 393
robots, reactive, 183
robots, with emotions, 174
Rogerian psychotherapy, 325, 326
role play, 92
Rosebud, 28ff
rule of goal generation, 240-241
rule of goal generation, as applied to values, 250-251

S

SAGE authoring language, 68, 69, 70, 71
SAGE computation module, 68
SAGE interface, 68, 72
SAGE (Story telling Agent Generation Environment), 61ff
sample store, 217
schizophrenia, 6, 24
science, 8, 14, 19, 21, 23
scripting language, 87, 104
scripts, 62, 65, 69, 292
secondary emotions, 178
self control (self-control), 174, 179, 180
self control, limitations of, 188
self monitoring, evaluation, 186
self, 27, 28, 30, 31, 32, 33, 34, 54, 55, 62, 63, 66, 75, 79, 83, 263-266, 269, 270-274, 277-287, 298
self-decision, 388
self-exploration, 36
self-expression, 34, 37
self-identity, 28, 54, 55
self-interest, 281, 283
self-observer, 101
self-preservation, 115, 120
self-representation, 86, 91
seminar coordinator, 116
seminar organizer, 115, 119, 128, 131, 132, 133
seminar-like events, 115, 120
seminars, 115, 118, 120, 128
semi-structured environment, 200, 201
sharable semantics, 310
shift internalization, 268
signification, 389, 391
SIM_AGENT toolkit, 192
Simulacra, 388
simulation of chemical controls, 188
simulations, 379, 381, 383, 384, 390
situated cognition, 398
slipnet, 119-122, 125, 128-131
social affordances, 424
social agent psychology, 298
social augmentation, 424
social behavior, 302, 304
social competence, 264, 265
social convention, 139
social dyad, 277, 290, 291
social emotions, 154, 155, 157
social engine, 139
social history, 293, 294
social insects, 200, 213
social intelligence, 86, 88, 100, 309

social interactions, 372
social protocol, 276
social psyche, 264
social relations, 65
social roles, 264, 276, 288, 292
social self, 264, 281, 287, 298
social space, 282, 286, 292, 293, 294, 295
social understanding, 421
social world, 264, 267, 275, 276, 281, 284, 286, 287, 290, 292-295, 298
socialization, 263, 298
socially intelligent agents (SIA), xvff, 417
socially situated intelligence (SSI), 115
socially situated, 113, 114, 115, 131, 133
society of mind, 199, 221
society, structure and functioning of, 139
soft interface, 66
software agent, 113, 114, 115, 133, 134, 135
software agents, reactive, 183
software robots, 379, 382
souls, 349
sources of controlled variation, 314
Space Madness, 330
sparse distributed memory, 115, 119, 123
spotlight controller, 124, 126, 127
SSS (Servo, Sybsumption, Symbolic), 219, 222
standard, 234
stigmergic, 405
stigmergy, 213
stories and memories, 34, 54
stories, 27ff, 61-65, 68, 69, 71, 72, 73, 75, 76, 80, 82
story space, 43, 48, 49
story telling interaction, 61, 68, 69
story telling (storytelling), 34, 35, 36, 38, 39, 44, 47, 48, 49, 51, 52, 54, 55, 61-69, 71, 73, 74, 79, 80
Storybook, 43, 44, 47, 49, 50, 51
story-evoking, 28, 34, 43, 55
storyteller(s), 36, 61, 62, 64-69, 71, 72, 74, 76, 80
strategic information, 266
strategic representation, 266, 267, 268, 271
strategic state, 265, 297
structurally coupled, 114
stuffed animals, 27, 28, 34, 35, 36, 39, 40, 41, 45, 47, 52, 53, 54
subsumption architecture, 219
supervenience, 167, 171
symbol grounding, 198
symbol manipulation in humans, 192
Sympathetic Sentience, 404ff
synthetic characters, 85, 86, 90, 91, 92, 96, 97, 100

T

3T architecture, 219
Tamagotchis, 379, 381, 383, 385, 386, 391
tangibility, 34, 54
tangible media, 28, 29, 30, 55
tangible user interface, 67
technological externalizations, 416
technological history, 415
teleo-reactive programs, 181
templates, 118
tertiary emotions in other animals, 180
tertiary emotions, 169, 174, 179, 180
testing theories, 193
therapy for robots, 193
therapy, 191
thought, 165
Time Traveller, 340
TinyMUCK, 332, 340
toy, 34, 35, 37, 38, 40-43, 45, 46, 47, 49, 51, 52, 53, 54
tracking memory, 118, 132
trait theories, 89, 101
troublemaker, 332
trust, 93
Turing machine, 324
Turing test, 324, 327, 328, 331, 342,

346, 347

U
UCPOP, 206, 218, 222, 223
unconscious processing, 172
unintelligent intelligence, 422
UNIX-based system, 115
user expectations, 323, 345, 346
user interface, 116
utility function, 204, 215

V
value, absolute meaning of, 248
value, and multifunctionality, 255-256
value, and norm, 251-252
value, and socialization, 254, 256
value, and terminal goal, 249-251
value, as a broken evaluation, 249, 255
value, cognitive function of, 254-256
value, definition of, 248
value, general features of, 252-253
value, introjection of, 254
value, relative meaning of, 248
value, social function of, 254
values, 267, 269, 272, 280, 283, 296, 297, 418
variation space, 312f
vehicles of meaning, 386
virtual agents, 165
virtual environments, 89, 90, 91
virtual machines, 172
virtual machines, causal powers of, 171
virtual machines, reality of, 172
Virtual Mattie, 132
virtual parallel machines, 375
Virtual Teletubbies, 220, 221
vision, 177
vision, in squirrels and birds, 165
Visitor-Hoster system, 115
visual appearance, 86, 91, 103, 104
visual arts, 395
visual authoring language, 62
visual representation, 97
VMattie, 132

W
wearable computer, 143
web pages, 85, 86, 87, 89, 90, 93, 95, 98, 107, 109, 110
what if? reasoning, 176, 178, 179, 184
wholism, 12
wholist, 330
working memory, 120, 121, 127
world model, 199, 201
Wrapping Knowledge Base (WKB), 316
Wrapping, 303, 315, 317

Z
zombies, 193

In the series ADVANCES IN CONSCIOUSNESS RESEARCH (AiCR) the following titles have been published thus far or are scheduled for publication:

1. GLOBUS, Gordon G.: *The Postmodern Brain*. 1995.
2. ELLIS, Ralph D.: *Questioning Consciousness. The interplay of imagery, cognition, and emotion in the human brain*. 1995.
3. JIBU, Mari and Kunio YASUE: *Quantum Brain Dynamics and Consciousness. An introduction*. 1995.
4. HARDCASTLE, Valerie Gray: *Locating Consciousness*. 1995.
5. STUBENBERG, Leopold: *Consciousness and Qualia*. 1998.
6. GENNARO, Rocco J.: *Consciousness and Self-Consciousness. A defense of the higher-order thought theory of consciousness*. 1996.
7. MAC CORMAC, Earl and Maxim I. STAMENOV (eds): *Fractals of Brain, Fractals of Mind. In search of a symmetry bond*. 1996.
8. GROSSENBACHER, Peter G. (ed.): *Finding Consciousness in the Brain. A neuro-cognitive approach*. n.y.p.
9. Ó NUALLÁIN, Seán, Paul MC KEVITT and Eoghan MAC AOGÁIN (eds): *Two Sciences of Mind. Readings in cognitive science and consciousness*. 1997.
10. NEWTON, Natika: *Foundations of Understanding*. 1996.
11. PYLKKÖ, Pauli: *The Aconceptual Mind. Heideggerian themes in holistic naturalism*. 1998.
12. STAMENOV, Maxim I. (ed.): *Language Structure, Discourse and the Access to Consciousness*. 1997.
13. VELMANS, Max (ed.): *Investigating Phenomenal Consciousness. Methodologies and Maps*. n.y.p.
14. SHEETS-JOHNSTONE, Maxine: *The Primacy of Movement*. 1999.
15. CHALLIS, Bradford H. and Boris M. VELICHKOVSKY (eds.): *Stratification in Cognition and Consciousness*. 1999.
16. ELLIS, Ralph D. and Natika NEWTON (eds.): *The Caldron of Consciousness. Motivation, affect and self-organization - An anthology*. n.y.p.
17. HUTTO, Daniel D.: *The Presence of Mind*. 1999.
18. PALMER, Gary B. and Debra J. OCCHI (eds.): *Languages of Sentiment. Cultural constructions of emotional substrates*. 1999.
19. DAUTENHAHN, Kerstin (ed.): *Human Cognition and Social Agent Technology*. 2000.
20. KUNZENDORF, Robert G. and Benjamin WALLACE (eds.): *Individual Differences in Conscious Experience*. 2000.
21. HUTTO, Daniel D.: *Beyond Physicalism*. n.y.p.
22. ROSSETTI, Yves and Antti REVONSUO (eds.): *Beyond Dissociation. Interaction between dissociated implicit and explicit processing*. n.y.p.
23. ZAHAVI, Dan (ed.): *Exploring the Self. Philosophical and psychopathological perspectives on self-experience*. n.y.p.

VILLANOVA UNIVERSITY
3 9346 00783077 3

NO LONGER PROPERTY OF
FALVEY MEMORIAL LIBRARY

Human Cognition and Social Agent Technology is written for readers who are curious about what human (social) cognition is, and whether and how advanced software programs or robots can become social agents. Topics addressed in 16 peer-reviewed chapters by researchers at the forefront of agent research include: Narrative intelligence and implementations of story-telling systems, socially situated avatars and 'conscious' software agents, cognitive architectures for socially intelligent agents, agents with emotions, design issues for interactive systems, artificial life agents, contributions to agent design from artistic practice, and a Cognitive Technology view on living with socially intelligent agents. The book addresses both software and robotic agents. On the one hand justice is done to the scientific and technical aspects, and on the other hand the reader will learn about pioneering technological developments which are necessary for a public discourse and critical evaluation on where social agent technology is leading us and how such a development can be shaped in order to meet the social, cultural and cognitive needs of humans.

The book is suitable for students, researchers, and everyone interested in this emerging and quickly growing field, it does not require any specialist background knowledge.

John Benjamins Publishing Company

ISBN 90 272 5139 8 (Eur) / 1 55619 435 8 (US)