Cartographies of the Mind

Studies in Brain and Mind

Volume 4

Series Editors

John W. Bickle, *University of Cincinnati, Cincinnati, Ohio*
Kenneth J. Sufka, *University of Mississippi, Oxford, Mississippi*

Cartographies of the Mind

Philosophy and Psychology in Intersection

by

Massimo Marraffa
Università Roma Tre, Rome, Italy

Mario De Caro
Università Roma Tre, Rome, Italy

and

Francesco Ferretti
Università Roma Tre, Rome, Italy

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-5443-2 (HB)
ISBN-13 978-1-4020-5443-3 (HB)
ISBN-10 1-4020-5444-0 (e-book)
ISBN-13 978-1-4020-5444-0 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

All Rights Reserved
© 2007 Springer
No part of this work may be reproduced, stored in a retrieval system, or transmitted
in any form or by any means, electronic, mechanical, photocopying, microfilming, recording
or otherwise, without written permission from the Publisher, with the exception
of any material supplied specifically for the purpose of being entered
and executed on a computer system, for exclusive use by the purchaser of the work.

Somebody who wants the truth
becomes a scientist; somebody
who wants to give free play to his
subjectivity may become a writer;
but what should somebody do
who wants something in
between?

Robert Musil

TABLE OF CONTENTS

Contributors xi

Preface xv

I. THE INTERPLAY OF LEVELS

1 Setting the stage: Persons, minds and brains 3
Massimo Marraffa

2 Computational explanation and mechanistic explanation of mind 23
Gualtiero Piccinini

3 Computationalism under attack 37
Roberto Cordeschi and Marcello Frixione

II. DIMENSIONS OF MIND

4 Vision science and the problem of perception 53
Alfredo Paternoster

5 Synaesthesia, functionalism and phenomenology 65
Fiona Macpherson

6 Integrating the philosophy and psychology of memory: Two case studies 81
John Sutton

7 Emotion and cognition: A new map of the terrain 93
Craig De Lancey

8 Categorization and concepts: A methodological framework 105
Cristina Meini and Alfredo Paternoster

9 Errors in deductive reasoning 117
Pierdaniele Giaretta and Paolo Cherubini

10 Language and comprehension processes 131
Elisabetta Gola

III. DIMENSIONS OF AGENCY

A. *Self-Knowledge*

11 The unconscious 147
Giovanni Jervis

12 Self-deception and hypothesis testing 159
Alfred R. Mele

13 Autonomous agency and social psychology 169
Eddy Nahmias

B. *Consciousness*

14 The cognitive role of phenomenal consciousness 189
Tiziana Zalla

15 The unity of consciousness: A cartography 201
Tim Bayne

16 Extended cognition and the unity of mind. Why we are not "spread into the world" 211
Michele Di Francesco

C. *Agency and the Self*

17 Extreme self-denial 229
Ralph Kennedy and George Graham

18 Empirical psychology, transcendental phenomenology, and the self 243
Stephen L. White

19 How to deal with the free will issue: The roles of conceptual analysis and empirical science 255
Mario De Caro

D. *Social Agency*

20 The beliefs of mute animals 271
Simone Gozzano

21 Naive psychology and simulations 283
Cristina Meini

22 The social mind 295
Francesco Ferretti

23 Social behaviors and brain interventions: New strategies for reductionists 309
Aaron Kostko and John Bickle

References 319

Index of names 363

Index of subjects 369

CONTRIBUTORS

Tim Bayne is Lecturer in Philosophy at St. Catherine's College, Oxford. His research interests include philosophical psychology and philosophical psychopathology. At present he is completing a book on the unity of consciousness.

John Bickle is Professor and Head of the Department of Philosophy and Professor in the Neuroscience Graduate Program at the University of Cincinnati. He is the author of *Psychoneural Reduction: The New Wave* (MIT Press 1998), *Philosophy and Neuroscience: A Ruthlessly Reductive Account* (Kluwer 2003), co-author (with Ronald Giere and Robert Mauldin) of *Understanding Scientific Reasoning*, 5th Edition (Thomson 2005), and editor of *The Oxford Handbook of Philosophy and Neuroscience* (Oxford UP, forthcoming). His research interests include the philosophy of neuroscience, scientific reductionism, and the cellular and molecular mechanisms of cognition and consciousness, on which he has published over forty papers and book chapters.

Paolo Cherubini is Professor of Psychology in the Department of Psychology at the Università di Milano-Bicocca. His main research interests are in the cognitive experimental psychology of reasoning, thought, and decision making.

Roberto Cordeschi is Professor of Philosophy at the Università di Roma "La Sapienza", where he teaches philosophy of science. He is the author of several publications in the history of cybernetics and in the epistemological issues of cognitive science and artificial intelligence, including *The Discovery of the Artificial* (Kluwer 2002).

Mario De Caro is Associate Professor in the Department of Philosophy at the Università Roma Tre; since 2000, he has also been teaching philosophy at Tufts University. Beside two books in Italian and several papers and book chapters in the areas of the philosophy of mind, philosophy of action and ethics, he is the editor of *Interpretations and Causes: New Perspectives on Donald Davidson's Philosophy* (Kluwer 1999), and the coeditor, with David Macarthur, of *Naturalism in Question* (Harvard UP 2004) and *Normativity and Nature* (Columbia UP, forthcoming).

Craig De Lancey is Assistant Professor of Philosophy at the State University of New York at Oswego. His publications include *Passionate Engines: What Emotions Reveal about Mind and Artificial Intelligence* (Oxford UP 2001).

Michele Di Francesco is Professor of Philosophy at the Università Vita-Salute S. Raffaele in Milan, where he teaches philosophy of mind and philosophy of cognitive sciences. His research interests include the ontology of mind and the philosophical consequences of cognitive science. In this connection he is dealing with issues concerning consciousness and the unity of the mind, mental causation, and emergentism. He is president of the Italian Society for Analytic Philosophy.

Francesco Ferretti is Lecturer in the Department of Philosophy at the Università Roma Tre, where he teaches philosophy and cognitive science. He has published a book (in Italian) on mental imagery and at present he is completing a book on language, mind and human nature.

Marcello Frixione is Associate Professor of Philosophy at the Università di Salerno, where he teaches philosophy of language and logic. His research interests are in the field of cognitive science, and include philosophical and epistemological issues, computation and cognition, knowledge representation in artificial intelligence and robotics.

Pierdaniele Giaretta is Professor of Philosophy at the Università di Verona, where he teaches philosophy of science and logic. His primary research interests for most of the last decade have been in Russell's logic and philosophy of logic, formal ontology, logic and reasoning. He has also worked on the role of ontology in knowledge representation, and the logical analysis of clinical diagnosis.

Elisabetta Gola is Lecturer in the Department of Philosophy at the Università di Cagliari, where she teaches philosophy of mind. Her research interests include artificial intelligence tecniques applied to natural language comprehension and philosophical issues related to non-literal uses of meaning, on which he has published two books (in Italian) and numerous papers and book chapters.

Simone Gozzano is Professor of Philosophy at the Università di L'Aquila, where he teaches philosophy of mind. His interests include intentionality (he wrote two books in Italian on the subject) and animal cognition (he edited two books on this topic). At present he is completing a book on mental causation.

George Graham is the A.C. Reid Professor of Philosophy and Adjunct Faculty Graduate Program in Neuroscience at Wake Forest University in Winston-Salem. His research interests include philosophy of mind, philosophical psychopathology, and the conceptual foundations of cognitive science. Among his recent publications

are the following: *Oxford Textbook in Philosophy and Psychiatry*, with K.W.M. Fulford and T. Thornton (Oxford UP 2006); and "Self-Ascription: Thought Insertion" in J. Radden (ed.), *The Philosophy of Psychiatry: A Companion* (Oxford UP 2004).

Giovanni Jervis is Professor of Dynamic Psychology at the Università di Roma "La Sapienza". A student of ethnologist Ernesto De Martino, his research has focused on social psychiatry (and psychology) and the foundations of psychoanalytic theories since the 1950s.

Ralph Kennedy is Associate Professor and Chair in the Department of Philosophy at Wake Forest University in Winston-Salem. He is currently working (with Michaelis Michael) on a paper on what it is to think of something as an object. His past papers include "How not to derive 'is' from 'could be': Professor Rowe on the ontological argument" (*Philosophical Studies* 1989) and, with Charles Chihara, "The Dutch book argument: its logical flaws, its subjective sources" (*Philosophical Studies* 1979).

Aaron Kostko is a Ph.D. student in the Department of Philosophy at the University of Cincinnati. His primary areas of specialization are philosophy of mind and philosophy of science, with particular emphasis on the emerging social neurosciences. His dissertation research attempts to provide an empirically informed account of the formation of self-conceptions and self-narratives, integrating findings from neuroscience and social psychology.

Fiona Macpherson is Lecturer in Philosophy at the University of Glasgow, where she established and directs the Centre for the Study of Perceptual Experience. She is spending the academic year 2005-2006 on secondment at the Centre for Consciousness in the Research School of Social Science at the Australian National University. She has published papers in the philosophy of mind in journals such as *Noûs*, *Philosophy and Phenomenological Research* and *Pacific Philosophical Quarterly.*

Massimo Marraffa is Lecturer in the Department of Philosophy at the Università Roma Tre, where he teaches philosophy and psychology. His research focuses primarily on issues in the philosophy of psychology and the foundations of cognitive science, on which he has published three books (in Italian) and several papers and book chapters.

Cristina Meini is Lecturer in the Department of Humanities at the University of Piemonte Orientale in Vercelli, where she teaches psychology. Her research focuses on the philosophy of cognitive processes. In particular, her interests include the ontogenesis and phylogeny of folk psychology, on which she has published two books (in Italian) and several papers.

Alfred R. Mele is the William H. and Lucyle T. Werkmeister Professor of Philosophy at Florida State University. He is the author of *Irrationality* (1987), *Springs of Action* (1992), *Autonomous Agents* (1995), *Motivation and Agency* (2003), and *Free Will and Luck* (2006), editor of *Philosophy of Action* (1997) and co-editor of *Mental Causation* (1993) and *The Oxford Handbook of Rationality* (2004), all published by Oxford University Press. He is also the author of *Self-Deception Unmasked* (Princeton UP 2001).

Eddy Nahmias is Assistant Professor in the Department of Philosophy and the Brains & Behavior program at Georgia State University. His research is in the philosophy of mind and moral psychology, focusing on questions about human agency: what it is, how it is possible, and how it accords with scientific accounts of human nature. Nahmias is currently writing a book manuscript, *Free Will and the Sciences of the Mind*.

Alfredo Paternoster is Associate Professor of Philosophy of Language at the Università di Sassari. His research interests include philosophy of language, philosophy of mind and foundations of cognitive science on which he has published two books (in Italian) and over 25 papers and book chapters.

Gualtiero Piccinini is Assistant Professor of Philosophy at the University of Missouri in St. Louis. He works primarily in philosophy of mind, with an eye to psychology, neuroscience, and computer science. His articles have been published in *Philosophy of Science*, *Australasian Journal of Philosophy*, *Philosophical Studies*, *Synthese*, *Canadian Journal of Philosophy*, *Studies in the History and Philosophy of Science*, *Journal of Consciousness Studies*, and *Minds and Machines*.

John Sutton is Associate Professor and Head of the Department of Philosophy at Macquarie University in Sydney. He works in the philosophy of cognitive science and the history of science, and is the author of *Philosophy and Memory Traces: Descartes to connectionism* (Cambridge UP 1998), and coeditor of *Descartes' Natural Philosophy* (Routledge 2000). His recent work is on skill memory, autobiographical memory, language and memory, dreaming, distributed cognition, and the extended mind hypothesis.

Stephen L. White is Associate Professor in the Department of Philosophy at Tufts University. He works in the philosophy of mind, epistemology, theory of action and ethics, and is the author of *The Unity of the Self* (MIT Press 1991) and *The Necessity of Phenomenology* (Oxford University Press, forthcoming).

Tiziana Zalla is a cognitive scientist at the French Centre National de la Recherche Scientifique (CNRS) presently working at the *Institut Jean-Nicod* in Paris. She worked at the *National Institute of Health* in Bethesda and at the *Institut de Sciences Cognitives* in Lyon. She has published many papers on consciousness, intentionality and knowledge of action in patients with schizophrenia and autism.

PREFACE

This book aims at exploring the potential for interaction between philosophy of mind (the area of philosophy that deals with our commonsense conception of mental matters—commonly called "folk psychology") and the science of psychology.

When we consider the relationship between these two domains of inquiry, we find a spectrum of positions. At one extreme, there is the idea that the investigation into the mental is the prerogative of either discipline. This perspective, sometimes termed "isolationism", can take two different forms.

According to *scientific* isolationism, the problems of philosophy of mind are either illusory or the prerogative of scientific psychology. In this perspective, the proper business of the philosophy of mind is, at most, the accurate re-description of the problems traditionally regarded as its area of expertise so that they can be handed over to empirical research.

By contrast, *philosophical* isolationism claims that philosophy of mind can proceed quite independently from any scientific enterprise: either because the very idea of a "science" of mind is seen as some sort of Rylean category mistake; or, less radically, because philosophical inquiry is conceived as having a "purely conceptual" or "transcendental" character, and hence it is constitutively autonomous from empirical research.

We believe that both forms of isolationism are to be rejected. Scientific isolationism is constantly at risk of loosing the mental as its own object of study, replacing it with objects that belong to different levels of analysis. By contrast, philosophical isolationism easily runs the risk of getting caught in a circle within a conceptual framework that is assumed to be necessary and universal, but manifestly rests on the dubious analytic/synthetic distinction.

Fortunately, there is a second, much more promising point of view—which can be called "interactionism"—, according to which philosophy of mind and scientific psychology should interact in the attempt to offer an integrated picture of the mental. In this perspective, contrary to philosophical isolationism, philosophy of mind is constrained by the findings of empirical research; but, contrary to scientific isolationism, it makes a non-replaceable contribution to the study of the mental by imposing on scientific psychology (through a methodology different from that of empirical investigation of the world) some crucial top-down constraints that derive from our folk psychological conceptual scheme.

In this perspective, the term "philosophy of psychology" is an appropriate label for the study of the interaction between an empirically-informed philosophy of mind and a philosophically-informed scientific psychology. This interaction consists in working back and forth between the ordinary image of ourselves as self-conscious, intentional, rational agents, and the scientific conception of ourselves as biochemically-implemented computational machines, by revising these two images wherever necessary so as to pursue the regulative ideal of a coherent self-conception.

The book comprises three parts. In the first part, "The interplay of levels", philosophy of psychology explores some foundational issues in scientific psychology. Here the focus is on the very possibility of a scientific psychology, with respect to both the legitimacy of its own level of analysis—the information-processing level—and the relationship that this entertains with, on the one hand, the lower level of the neurosciences, and, on the other hand, the higher level of the philosophical reconstruction of our folk psychological conceptual scheme.

The second part of the book, "Dimensions of mind", gets inside psychology, and the interactive approach is applied with the aim of clarifying issues and debates concerning some classical mental phenomena (vision, synaesthesia, memory, emotions, concepts, reasoning and language). Finally, in the third part, "Dimensions of agency", this approach comes to grips with some thorny issues, which traditionally are considered impervious to the projects of naturalization, and hence are often paraded as evidence in favor of philosophical isolationism (self-knowledge, consciousness, the self, free will and social agency). Unsurprisingly, the authors of the essays collected in this section of the book differ in their views about how harmonious the interplay of philosophical analysis and empirical investigation is likely to be with regard to these issues; everybody, however, agrees about its fruitfulness.

The Editors

Rome, November 2006

PART I THE INTERPLAY OF LEVELS

CHAPTER 1

SETTING THE STAGE: PERSONS, MINDS AND BRAINS

Massimo Marraffa

Over the last thirty years, the philosophy of science has become increasingly "local". Its focus has shifted from the general features of scientific enterprise to the concepts, theories, and practices of particular disciplines. Philosophy of neuroscience, philosophy of psychology, and philosophy of cognitive science are three results of this growing specialization.[1]

This chapter is a very short introduction to the philosophy of *cognitive* psychology, especially in its *computational* incarnation. Cognitive psychology investigates complex organisms at the *information-processing* level of analysis, and it can be defined a peculiar level in the sense that it is suspended between two worlds. On the one hand, there is the ordinary image of ourselves as *persons*, namely as self-conscious, intentional, rational agents. On the other hand, we have the subpersonal sphere of the cerebral events, as investigated by neuroscience. Therefore, one of the main tasks for the philosopher of psychology is to unravel this peculiarity, trying to shed some light upon the relations between these different ways of describing ourselves.

The following pages are dedicated to some classical attempts to accomplish this task. In the course of doing so, we shall draw a very quick sketch of the rise and development of cognitive psychology and cognitive science, setting the scene for the other chapters of this book.

1. FROM FOLK PSYCHOLOGY TO COGNITIVE SCIENCE

1.1 The form and the status of folk psychology

Folk psychology as a theory. To navigate through the social world normal adults advert to a spontaneous capacity to "mentalize" or "mindread", that is, to describe, explain, and predict their own and other people's behavior on the basis of mental state attributions.[2]

According to the so-called "theory theory", mindreading rests on a theory, or rather a proto-theory, often called "folk psychology". This is a theory[3] in the sense of being an integrated and coherent body of knowledge which organizes the multiform sphere of the mental essentially through two categories: qualia and

M. Marraffa, M. De Caro and F. Ferretti (eds.), Cartographies of the Mind, 3–22.
© 2007 *Springer*.

intentional states. The former are the experiential or introspectible properties of mental states. Their essence seems to consist in their being captured from a subjective or first-person point of view—there is something that it is like to perceive a shade of red or to regret that Brutus killed Caesar. As a whole, these mental entities define the domain of *phenomenal consciousness*.[4]

In contrast, intentional states are states (such as believing, desiring, regretting, etc.) which have "direction toward an object" or "reference to a content".[5] If I believe that Brutus killed Caesar, my belief is directed toward an object or refers to a content, that is, what the sentence "Brutus killed Caesar" expresses. Intentional states are often termed "propositional attitudes" since—as the example shows—in ascribing them to a subject, we use sentences of the form "*S* believes (or desires, etc.) that *p*", where the proposition *p* expresses the content of the subject *S*'s mental state.

In any intentional state, the objects on which the state is directed are presented in a certain way, namely it has a *representational* character. When I believe that London is north of Paris, I represent a state of affairs in the form of a particular spatial relation between two objects. This point is often made by saying that intentional states are *semantically evaluable*, that is, they can be true or false—my belief that London is north of Paris is true if there is a fact in the world that makes it true.[6]

Compatibilism vs. Eliminativism. Social psychologists have investigated mindreading since at least the 1940s. In Heider and Simmel's classic studies, subjects were presented with geometric shapes that were animated as if moving around in relation to each other. When asked to report what they saw, the subjects almost invariably treated these figures as intentional agents with motives and purposes, suggesting the existence of a universal and largely automatic capacity for mentalistic attribution.[7]

Pursuing this line of research would lead to Fritz Heider's *The Psychology of Interpersonal Relations* (1958), a seminal book that is the main historical referent of the inquiry into folk psychology. In particular, it played a central role in the origination and definition of *attribution theory*, a field of social psychology that investigates the mechanisms underlying ordinary explanations of our own and other people's behavior.

Attribution theory is an offspring of Heider's visionary work, but a quite different way of approaching folk psychology. Heider takes folk psychology in its real value of knowledge, arguing that "scientific psychology has a good deal to learn from common-sense psychology".[8] In contrast, most research on causal attribution is true to behaviorism's methodological lesson and focuses on folk psychology's naivetes.[9]

The contrast between these two attitudes toward the explanatory adequacy of folk psychology has shaped the philosophical debate on the fate of the ordinary image of ourselves in light of the tumultuous development of cognitive science. On this matter the basic issue is: will the theoretical entities invoked in folk psychology be a part of the ontology of a serious scientific psychology? And the answers range

from Jerry Fodor's "definitely yes", based on the idea that propositional attitudes are the bedrock of a scientifically adequate psychology; to Stephen Stich's "possibly not", motivated by doubts about the folk concept of belief raised, *inter alia*, just by attribution theory;[10] to Paul Churchland's "absolutely not", based on the idea that the deliverance from folk concepts is the condition of psychology's being reducible to neuroscience, and hence having a scientific nature.

These two perspectives on the status of folk psychology—the former "compatibilist", the latter "eliminativist"—are the coordinates that help us to navigate through the complex conceptual landscape of the cognitive revolution. As we shall see, the rise of cognitive psychology is the result of the rejection of the behaviorist eliminativism (subsection 1.2) in favor of a compatibilist project which represents a sort of "experimental mentalism"[11] (subsection 2.1). Nevertheless, the eliminativist ghost has continued to haunt cognitive psychology, taking on always new forms (subsection 2.2).

1.2 The rise and fall of behaviorism

Psychology as phenomenology. Both the classical empiricist and the classical rationalist pictures of introspective self-knowledge (or, in up-to-date terms, "first-person mindreading") have granted it a special epistemic authority. According to Descartes, for example, the subject is transparent to itself, and the reflective awareness (*conscientia*) the mind has of its own contents provides knowledge enjoying a special kind of certainty, which contrasts with our knowledge of the physical world: the judgments about our current mental states and processes are infallible or, at least, incorrigible.

In light of this traditional optimism about self-knowledge, it is not at all surprising that in the late nineteenth and early twentieth centuries scientific psychology is predominantly a psychology of introspective consciousness.[12] Pursuing the project to make introspection a rigorous method of inquiry, which would upgrade psychology to the status of the other natural sciences, early experimental psychologists meticulously probed the contents of consciousness in an effort to offer a full description of the mental landscape as it appears to the subject. In short, this psychology was "a kind of phenomenological investigation of subjective self-awareness".[13]

Eliminative behaviorism. By virtue of the mentalistic idiom, these introspectionist psychologists would not have trouble talking to "poets, critics, historians, economists, and indeed with their own grandmothers. The nonspecialist reader in 1910 would be in equally familiar territory in William James's *Principles of Psychology* and in the novels of James's brother Henry".[14] John Watson's brand of behaviorism put an end to the good relationship between scientific psychology and folk psychology, urging to abandon the introspectionist attempts to make consciousness a subject of experimental investigation. A psychology aspiring to scientific respectability had to rely instead on publicly observable data, that is,

patterns of responses (overt behavior) to stimuli (physical events in the environment). The outcome was an extremely austere conception of psychological explanation: the psychologist, equipped with nothing but Pavlov's conditioning and Thorndike's law of effect (precursor of Skinner's operant conditioning), had to chart associative connections between classes of environmental inputs (or histories of exposure to environmental inputs) and classes of behavioral outputs. What occurred in the "head", between input and output, was a topic for physiology (the ultimate behavioral science). The organism was regarded as a "black box".

Insofar as behaviorism removes the inner states and processes from psychology's explanation and ontology, it can be considered a variant of the doctrine of *eliminativism*.[15] In its strongest form, eliminativism predicts that part or all of our folk psychological conceptual apparatus will vanish into thin air just as it happened in the past, when scientific progress led to drop the folk theory of witchcraft or the protoscientific theories of phlogiston and caloric fluid. This prediction rests on an argument which moves from the premise of considering folk psychology as a massively defective theory to the conclusion that—just as witches, phlogiston and caloric fluid—folk psychological entities do not exist. (Sometimes this negative ontological conclusion is set by the weaker conclusion that folk psychological entities will not be part of the ontology of a mature science.) The behaviorist version of eliminativism predicts that the scientific theory which replaces the seriously mistaken folk psychological theory will be couched in the vocabulary of physical behavior.

Eliminative behaviorism is a recurrent theme in the writings of Watson and Skinner, although in some passages they waver between an eliminative interpretation of behaviorism—an ontological and explanatory thesis: mental entities do not exist and hence the explanation of animal behavior will be *non*-mentalistic—and other two interpretations: (i) the *methodological* claim that mental entities exist but are irrelevant to the scientific study of animal behavior; and (ii) the *semantical* claim—known as "analytic" or "logical" behaviorism—that statements containing psychological terms are translatable into statements containing just terms referring to physical behavior. This is a *reductive* program: mental entities are not eliminated, but rather identified with dispositions to behave in certain ways under certain circumstances.[16]

A point is well worth emphasis. As Larry Hauser rightly says, "although behaviorism as an avowed movement may have few remaining advocates", some of its "metaphysical and methodological challenges" are still very much alive.[17] First and foremost, the fundamental objection that Skinner had to the mentalistic explanation in psychology, namely the *homunculus fallacy*, is a vital constraint on any serious mentalistic psychology. That is, a plausible theory of cognition must avoid the infinite regress triggered by the attempt to explain a cognitive capacity by tacitly positing an internal agent with that very capacity.[18]

Cognitive maps and syntactic structures. Since the 1930s and 1940s the increasing perception of the limits of the S(timulus)-R(esponse) explanation makes behaviorism evolve towards what would be, since the 1960s, cognitive psychology.

A landmark in this evolution is the classical series of rat experiments in the Berkeley laboratory of Edward C. Tolman. These experiments demonstrated that the maze-navigation behavior of rats could not be explained in terms of S-R mechanisms, leading Tolman to suggest that the animals were building up complex representational states, or "cognitive maps", which helped them locate reinforcers.[19] These results were pointing in the same direction as Kenneth Craik's suggestion that the mind does not work directly on reality, but rather on "small-scale models" of it.[20]

Some ingenious attempts to refine the S-R schema were made to account for Tolman's experimental results without his troublesome mentalistic concessions.[21] However, such a schema turned out to be totally powerless when the focus shifted from maze navigation behavior in rats to verbal behavior in human beings. Thus, it is hardly surprising that one of the main factors of the transition from behaviorism to cognitivism was the impetuous development, since the late 1950s, of a mentalistic theory of language, namely Noam Chomsky's generative linguistics.[22]

Over the course of his trenchant criticism of empiricist theories of linguistic learning, Chomsky put forward an argument that would become one of the tools of the cognitivist trade: the *poverty of the stimulus argument*.[23] Let's examine the input and the output of the process of first-language acquisition. A large amount of empirical evidence attests to a gap between the learning target achieved by the child (its mature linguistic competence) and the "primary linguistic data" (the child's observations of utterances produced by adult members of its speech community). In other terms, the output contains more information than was present in the input. This extra information can be nothing but a contribution made by the human learner, that is, the innate cognition of certain facts about universal constraints on possible human languages (the so-called "Universal Grammar").

1.3 Inside the black box: The vicissitudes of information

Biological information processing. The scope of the Chomskian argument goes far beyond the case of language acquisition. And it is not an overstatement to claim that "Modern Cognitivism starts with the use of poverty of the stimulus arguments".[24] If it turns out that there is more information in the response than there is in the stimulus that prompts the response, we must assume the intervention of some kind of inner processing of the stimulus. This work that the organism does is an *unobservable* cause that the cognitivist infers from behavior. And this is epistemologically correct, since postulating unobservables such as electrons and genes is the standard practice in science.

Therefore, cognitive psychology can be defined as the science that investigates the processing of information in the head, that is, "all the processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used".[25] Instead of the behaviorist "empty organism", cognitivists reintroduce the mind construed as an "information processor" intervening between the impingements on sensory organs and the behavioral response.[26] The input

information is codified in the mind, thus becoming inner objects—*mental representations*—that can undergo various types of processing. In particular, these objects can be *transformed*, which means that our representation of reality is not the product of a passive assimilation of physical environment, but an active construction that can involve both a *reduction* and an *integration*. Biological information processing is capacity-limited and hence necessarily selective. We can attend to a relatively small number of stimuli, and a still smaller amount of them can be recalled. Hence it is possible that part of the input information gets lost, and then a reduction takes place. Alternatively, sensory input may be integrated, enriched, and it is in such a case that some well-known poverty of the stimulus arguments concerning perception and memory make their appearance.

Perceptual constancies are a case in point. In the case of size constancy, for instance, the visual system takes account of the perceived distance of objects and scales perceptual size up accordingly. Therefore, in this case as in that of linguistic acquisition, there is more information in the perceptual response than there is in the proximal stimulus, and this extra information can be nothing but a contribution made by the perceiving organism.

Perceptual integration had attracted psychologists' attention well before the rise of cognitivism. Most notably, Hermann von Helmholtz considered perceptual processes as *unconscious inferences*, which take specifications of proximal stimulations as premises and yield hypotheses about their distal causes as conclusions. This *constructive* conception of perception has been named "Establishment View",[27] and, indeed, most of the work on vision that cognitive scientists have done since 1970s rests on this approach. In this lapse of time, however, constructivism did not go unchallenged. The advocates of J.J. Gibson's *ecological optics* have contended that "the visual system, far from reconstructing or inferring, merely extracts, picks out, the information present in the stimulation, 'attuning itself' to the relevant information structures".[28] And we shall see below (subsection 3.2) that Gibson is the main source of inspiration for a recent theory of cognition known as "active externalism".

Computational functionalism. According to a largely dominant interpretation, the processes of transformation, storage, recovery and use of information are *computations*, namely rule-governed sequences of operations upon data structures (mental representations), which mediate the organism's behavioral responses to perceptual stimuli.

The notion of computation here presupposed goes back to Alan Turing's work. His "Turing machines" are *abstract* computers since their characterization does not take into account constraints that are essential in planning a real computer (e.g., memory space and computing time), and above all in that they are defined without any reference to their physical makeup (i.e. the type of hardware that realizes them). In fact, Turing machine states are fully definable in terms of (i) the machine's input, (ii) the output of the machine given its state and that input, and (iii) the next state of the machine given the current state. That is, the states are

functionally defined since all that matters to what they are, is what the machine *does*, rather than its physical realization.

Now, if cognitive processes are computations, they also must be functionally individuated, that is, individuated by the causal role (or function) they play in the cognitive system of which they are a part, independently from how such a role is physically (or, better, neurologically) realized. This thesis on the essence of cognition is known as "computational functionalism".

Insofar as cognitive psychology subscribes to computational functionalism, it contributes to cognitive science, namely the project of interdisciplinary study of natural and artificial intelligence that begins its maturation in the late 1950s and reaches a stable intellectual and institutional set-up in the early 1980s.[29]

One point is worth emphasizing. Cognitive science is the study of cognition as information processing by a natural or artificial computer, but research in cognitive science is typically about a specific type of computer: for instance, computational psychology investigates the *biological* computer, whereas artificial intelligence (AI) explores the *artificial* one. Therefore, cognitive science is not a discipline, but rather a "doctrine" that has oriented and is orienting inquiries in a number of disciplines[30]—some descriptive and empirical (e.g., cognitive psychology, linguistics and, more recently, neuroscience), some speculative and foundational (e.g., philosophy), and some both speculative and applied (e.g., AI).[31]

David Marr's tripartite model of explanation. Computational functionalism underlies Marr's deeply influential analysis of how different levels of explanation can be integrated to understand a cognitive phenomenon.[32] This analysis can be regarded as "the first full-blown version of computationalism".[33]

After attempting to elucidate how the brain performs cognitive tasks by starting with the response patterns of individual neurons (e.g., Hubel and Wiesel's "on-centered" and "off-centered" cells), Marr realized that discovering such patterns is only a *description* of what is happening in the brain, not an *explanation* of how it discharges its tasks. Consequently, he concluded that a computational account of a cognitive phenomenon needs to integrate the level of analysis of the "wetware" with other two levels of analysis.

At the most abstract level of explanation is the "computational theory", where we specify what a system is doing and why. In Marr's theory of vision, for example, the function of the visual system is to construct, on the basis of inputs to the photoreceptors, a 3-D object-centred shape representation. At this level, psychological functions are characterized only in terms of their input data, the final output, and the goal of the computation, in ways that are neutral on the mechanism. Between the computational theory level and the level of "implementation" (as Marr terms the level of analysis of the wetware) is the "algorithmic" level. This level—which is the one specific to psychology—concerns the cognitive mechanism (the algorithm) that performs the function described at the level of the computational theory. For example, Marr outlines at the algorithmic level the intermediate representations between the retinal image and the final output (primal sketch and 2½-D sketch), and suggests some of the subsystems that compute them.[34]

2. FOLK PSYCHOLOGY AND COMPUTATIONAL PSYCHOLOGY

2.1 Folk psychological computationalism

Mind as a syntax-driven machine. What kind of relation is there between the computational states and processes postulated at the algorithmic level and the folk psychological mental states and processes? According to Fodor, the relation is one of legitimation or grounding for our folk psychological explanatory practice: "One can say in a phrase what it is that computational psychology has been proving so successful at: viz. *the vindication of generalizations about propositional attitudes*, specifically, of the more or less commonsense sorts of generalizations about propositional attitudes". Therefore, "[w]hat a computational theory does is to make clear the mechanism of intentional causation; to show how it is (nomologically) possible that purely computational—indeed, purely physical—systems should act out of their beliefs and desires".[35]

There are two dimensions to the problem of making clear "the mechanism of intentional causation", of showing how it is possible that purely physical systems should act out of their propositional attitudes.

The first problem concerns the nature of the intentional mental *states*. They are both *semantically evaluable* and *causally efficacious*, two properties that apparently never occur together elsewhere. This putative uniqueness has fed many perplexities about the perspectives of a physicalist explication of intentional states. For many philosophers they still remain, in Quine's famous phrase, "creatures of darkness".[36]

Actually, there is something else that is both semantically evaluable and causally efficacious: *symbols*. They can be about things (e.g., the word "cat" refers to cats); and they are physically instantiated or tokened, which makes them causally efficacious (the word "cat" consists of, e.g., ink on paper). Hence there is an analogy between thoughts and symbols, and "the history of philosophical and psychological theorizing about the mind consists largely of attempts to exploit it by deriving the causal/semantic properties of the former from the causal/semantic properties of the latter".[37] Fodor's Representational Theory of Mind (RTM) is the most recent heir to this tradition, claiming that intentional states are relations between an agent and mental representations regarded as symbols of a *Language of Thought* (LoT). This is a formal language akin to the first-order predicate calculus.

The second problem concerns the mechanics of thinking over time. The folk psychological laws that govern intentional mental *processes* subsume *causal* interactions among intentional states preserving their *semantic coherence*. For example, reasoning (the mental process par excellence) is a causal sequence of intentional states that tends to preserve their semantic (rational, epistemic) properties. But what if not an inner interpreter might be sensitive to such properties? Here RTM is at risk of the above-mentioned homunculus fallacy. Accordingly, a *mechanical explanation of rationality*—that is, the proof that a purely physical mechanism can implement causal interactions among intentional states preserving

their semantic coherence—needs a strategy to prevent the regress of inner interpreters.

This strategy, Fodor suggests, consists of combining RTM with the Computational Theory of Mind (CTM), namely the hypothesis that intentional mental processes are causal sequences of symbol transformations driven by rules that are sensitive to the syntactic form of the symbols and not to their content.

At the foundations of CTM there are the methods of proof theory and the Turing machines.[38]

The proof-theoretic approach in logic has showed us how to link up semantics to syntax. For any formalizable system of symbols, it is possible to specify a set of formal derivation rules which, albeit sensitive only to the syntactic form of symbols, allows us to make all and only the semantically valid inferences. In this way, certain semantic relations between symbols are "mimicked" by their purely syntactic ones.

The relevance of this result cannot be exaggerated. According to Fodor and Pylyshyn, the "classical" cognitive science can be described as "an extended attempt to apply the methods of proof theory to the modeling of thought (and, similarly, of whatever other mental processes are plausibly viewed as involving inferences; pre-eminently learning and perception)".[39] Accordingly, the hope is that "syntactic analogues can be constructed for non-demonstrative inferences (or informal, commonsense reasoning) in something like the way that proof theory has provided syntactic analogues for validity".[40]

Formalization suggests a strategy to bridge the gap between semantics and causal efficacy that blocks the mechanization of the semantic coherence of thought: in fact, given the connection that formalization makes between semantics and syntax, if a link was set up also between syntax and causal efficacy, then it would be possible to connect semantics with causation *via* syntax. Here is where Turing's theory of computability comes into play.

Any formalizable process can be characterized in terms of effectively computable functions (i.e. functions for which an algorithm can be given). As stated by the "Church-Turing thesis", all the effectively computable functions can be carried out by a Turing machine (assuming that both the tape and time are infinite). Since any Turing machine can be implemented by a physical mechanism (e.g., a digital computer), it follows that, for any finite formal system, it is possible to devise a machine which is able to automate the inferences of that system. Because certain of the semantic relations among the symbols in a formal system can be "mimicked" by their syntactic relations, and because such a system can be implemented by a computer, it follows that it is possible to construct a machine driven by syntax whose state transitions satisfy semantic criteria of coherence. Because digital computers are purely physical systems, this shows us that it is possible for a purely physical system to make inferences which respect the semantics of the symbols without invoking a question-begging homunculus.[41]

According to the Representational and Computational Theory of Mind (RCTM), the mind is a particular kind of computer, and the causal interactions among intentional states are implemented by computations on the syntactic

properties of LoT symbols, which are physically tokened in the brain like data structures in a computer. LoT is a formal system, and hence its rules preserve the semantic properties of the symbols. Minds are, in Dennett's oft-cited phrase, "syntactic engines that can mimic the competence of semantic engines".[42]

In RCTM the propositional attitude relations in RTM are identified with the computational relations in CTM. Each propositional attitude is identified with a characteristic computational/functional role played by the LoT sentence that it is the content of that kind of attitude. For example, a LoT sentence *p* might be the content of a *belief* since "it is characteristically the output of perceptual systems and input to an inferential system that interacts decision-theoretically with desires to produce further sentences or action commands".[43] Or equivalently, to believe that *p* is for *p* to be available to one set of computations, whereas to desire, to regret, to hope that *p* is for *p* to be available to other sets of computations.

Cognitive psychology as anti-phenomenology. Fodorean mentalism is not introspectionist mentalism in a new guise. As we have seen, the mind that Fodor takes as the subject of cognitive psychology is not the introspective consciousness, but a kind of formalization of the psychology of propositional attitudes. The propositional attitude states can occur both in explicit, conscious judgments and in mental states that the agent could not possibly introspect, even in principle. This presupposes that consciousness and intentionality can be studied in the absence of one another, an approach to mentality that would not have been possible in the pre-Freudian conceptual universe, where consciousness and intentionality were intrinsically linked. However, as Fodor reminds us, "Freud changed all that. He made it seem plausible that explaining behavior might require the postulation of intentional but unconscious states. Over the last century, and most especially in Chomskian linguistics and in cognitive psychology, Freud's idea appears to have been amply vindicated".[44]

Actually, on this matter one can be more radical than Fodor, claiming that cognitive psychology has not simply vindicated Freud, but has gone far beyond. In fact, the Freudian concept of the unconscious is parasitic to a concept of consciousness idealistically taken as "a primary quality of the mind",[45] whereas cognitive psychology has given rise to "a reinforcing overturning of traditional psychodynamic questions", and starts with asking how consciousness, rather than the unconscious, is possible.[46] In this way, cognitive psychology amends Freud in view of Darwin. That is, it follows Darwin's anti-idealistic methodological lesson and proceeds *bottom-up*, attempting to explain how the complex psychological functions underlying first-person awareness evolve from the more basic ones.[47] This attempt does not appeal to our introspective self-knowledge, but all those disciplines—first and foremost developmental psychology—that investigate the gradual construction of self-awareness. In other words, cognitive psychologists see the conscious subjective experience as "an advanced or derived mental phenomenon, not the foundation of all intentionality, all mentality";[48] or, in more Continental terms, cognitive psychology is an *anti-phenomenology*, that is, a critique of the subject, of its alleged "givenness".[49]

In the next section we shall see how cognitive psychology has addressed its critical potential not only against our phenomenological intuitions about consciousness and self-consciousness, but also against its own intentional grounds, thus opening the door to new behavioristic and eliminativistic objections.

2.2 Behavioristic and eliminativistic challenges to RCTM

Anti-introspectionism, externalism, and the syntactic theory of mind. The compatibilist view of the interface between propositional attitude psychology and scientific psychology takes the former as a good working hypothesis on the overall computational organization of the human mind. Noteworthy work in cognitive science has assumed that the folk account of the architecture of the mind is largely correct, though it is far from complete. However, there are also findings and theories which seem to suggest that our cognitive system is organized along lines quite different from those theorized by folk psychology. Here is a classic example.

In 1977, after reviewing the experimental social psychology literature on dissonance and self-attribution, Richard E. Nisbett and Timothy D. Wilson concluded that the reports about the causes of our behavior are not reconstructions of real mental states and processes, due to a direct introspective awareness, but rather a "confabulatory" activity, originated by the employment of "a priori causal theories".[50] In this perspective, introspection becomes a form of *self-deception*.[51]

These ideas have been hugely influential. In developmental psychology and cognitive psychiatry, the hypothesis that behind the illusion of a direct introspective access there is an inferential activity based on socially shared explanatory theories has been developed in the framework of the theory-theory approach to the inquiry into the cognitive mechanisms underlying mindreading. Here "theory" refers to a tacit knowledge structure, a body of mentally represented information driving the cognitive machinery underlying mentalization.[52] For most advocates of this approach, this theory underlies both self-attribution and hetero-attribution of mental states. Therefore, "even though we seem to perceive our own mental states directly, this direct perception is an illusion. In fact, our knowledge of ourselves, like our knowledge of others, is the result of a theory".[53]

Neuropsychology is another research area that abounds with phenomena undermining the reliability of introspective consciousness. Consider, for example, the "split-brain" syndrome.[54] Split-brain patients are patients whose corpus callosum has been severed. As a result, the hemispheres of their brains can no longer communicate with one another, giving rise to a complex array of deficits. Suppose, for example, that the command "Walk" is flashed to the right hemisphere of a split-brain subject: "the patient typically stands up from the chair and begins to take leave from the testing van. When asked where she is going, she (the left side of the brain) says, 'I'm going into the house to get a Coke'".[55] A possible explanation of this pattern of behavior is that the right hemisphere responds to the command by making inferences that the subject cannot introspect or report, whereas the left hemisphere

"interprets" the right hemisphere's response and tells an implausible story unconnected with the command.

We find a very similar hypothesis about the cognitive mechanisms underlying confabulation in Wilson (1985). He hypothesizes two relatively independent cognitive systems: an unconscious system underlying nonverbal behavior, and a largely conscious system, whose function is to attempt to verbalize, explain and communicate what is occurring in the unconscious system. The latter takes information from the former as input and makes inferences based on repertories of rationalizations afforded by theories about the self and the situation.

Reflecting on Wilson's hypothesis, Stich has highlighted its critical potential against the folk concept of belief. A fundamental tenet of folk psychology is that our cognitive system is so organized that the very same state which underlies the sincere assertion of "p" also may lead to a variety of nonverbal behaviors. But from Wilson's dual system hypothesis follows that this principle is radically wrong, and "in those cases when the verbal subsystem leads us to say 'p' and our nonverbal subsystem leads us to behave as though we believed some incompatible proposition, there will simply be no saying which we believe".[56] Therefore, Stich concludes, Wilson's model shows that the tenability of the folk conception of mental architecture, the legitimacy of taking it as the ground on which to build a scientific theory of the mind, "is very much an open empirical question".[57]

Stich (1983) combines these doubts about the sorts of states and mechanisms that folk psychology invokes with another line of eliminativist argumentation, focused on folk psychology's reliance on semantic content.

Earlier we saw that Fodor's argument for a scientific intentional psychology rests on a "correlation thesis", according to which differences in content are mirrored by differences in syntax.[58] It is thanks to this correlation that the semantic properties of mental states are causally implicated in the production of behavior. The thesis, however, seems to be false: the well-known Putnam's and Burge's arguments for *semantic externalism* seem to demonstrate that the ordinary semantic properties ("wide content" properties) of mental states do not supervene on their formal properties.[59] Hence Fodor cannot "have it both ways": he cannot endorse both an *individualistic* methodology (i.e. cognitive psychology should be restricted to quantifying over the formal properties of mental states) and the *scientific intentional realism* (i.e. the intentional properties of mental states, properties that are *not* formal, are and will be part of the ontology of the cognitive psychology).

Assuming that scientific psychology must be individualistic, a way out from this impasse is to deny that intentional properties have any legitimate role in scientific psychological explanation. Stich's "syntactic theory of mind" takes this eliminativist option, and argues that cognitive psychology should recast its theories and explanations in a way that does not appeal to the wide content properties of mental states, but only to their individualistic, formal properties.[60]

Externalism strikes again. RCTM is unquestionably the most powerful systematization of computational functionalism. It holds a pivotal position in contemporary philosophy of psychology because it was the first major synthesis of

functionalist philosophy of mind with the cognitive revolution in psychology, and with the first generation of AI. Over the last two decades, however, this theory has been under attack, mostly owing to the expansion of cognitive science in two directions: "vertically into the brain and horizontally into the environment".[61] The force propelling these downward and outward developments is the pressure put on the individualist, modular, computational and representational conception of the mind by neurosciences, neoconnectionist cognitive modeling, dynamic approaches to cognition, artificial life, real-world robotics, and other research programs sometimes grouped under the heading of "non- or post-classical" cognitive science.

The current debate on the conceptual foundations of cognitive science shows a range of positions which are characterized by the more or less radical attitude towards the implications of the post-classical body of work. At one end of the spectrum there is the claim that RCTM is "by far the best theory of cognition that we've got",[62] and the post-classical research programs are much ado about nothing. At the other end of the spectrum there is a view of the post-classical body of research as an exercise of extraordinary science, which preludes to the establishment of a new paradigm.[63] Then in between these two poles is a "revisionist" perspective, which accepts some critical requirements of the post-classical research programs—first and foremost the deep dissatisfaction with the antibiologism and individualism of RCTM—and uses them as guidelines to reconstruct the conceptual bases of cognitive science.

Andy Clark is a leading advocate of revisionism. He believes that RCTM can be reconstructed making due allowances for "the environmental embedded, corporeally embodied, and neurally 'embrained' character of natural cognition",[64] but without collapsing into the anti-representationalism characteristic of the most radical readings of post-classical cognitive science. Accordingly, Clark pursues the metamorphosis of RCTM into just one component in a three-tiered explanatory strategy:

(i) a *dynamicist* account of the gross behavior of the agent-environment system;[65]

(ii) a *mechanistic* account,[66] describing how the components of the agent-environment system interact to produce the collective properties described in (i);

(iii) a *representational* and *computational* account of the components identified in (ii).[67]

Clark calls this tripartite explanatory strategy "minimal representationalism", and puts it into a wider theoretical framework: the "active externalism".[68] Unlike the above-mentioned *semantic* externalism, where the mental contents of a subject depend on aspects of the environment which are clearly external to the subject's cognitive processes, active externalism asserts that the environment can play an active role in constituting and driving cognitive processes. In the wake of Gibson, this environment is viewed as a complex of "affordances", which brings to the

formation of internal states that describe partial aspects of the world and prescribe possible actions.[69] These are "action-oriented" representations which, unlike LoT symbols, are *personal* (in that they are related to the agent's needs and the skills that it has), *local* (in that they relate to the circumstances currently surrounding the agent) and *computationally cheap* (compared with Marr's rich inner models of the visual scene).

Clark's active externalism confirms a point we made earlier, namely the relevance to the present day of some behavioristic metaphysical and methodological challenges. In fact, insofar as "emphasis on the outward or behavioral aspects of thought or intelligence—and attendant de-emphasis of inward experiential or inner procedural aspects—is the hallmark of behaviorism",[70] active externalism *is* behavioristic.

Eliminative connectionism. Clark's revision of RCTM follows the anti-individualistic guidelines that characterize the body of research on situated and embodied cognition. Now we turn to another revision, which reflects the movement downwards, into the brain, arising from the connectionist cognitive modeling and computational neuroscience.

During the 1970s the functionalist approach inclined some scientifically-minded philosophers to view computational psychology as radically autonomous from neuroscience. For example, in *Special Sciences* Fodor draws a principled argument for a very strong autonomy of psychology from a combination of functionalism, multiple realizability thesis, and token-identity theory.[71] By the late 1970s, however, "some philosophers were objecting to the divorce of cognitive science from neuroscience, Paul M. and Patricia S. Churchland foremost amongst them. They tended to continue to endorse a version of the identity theory and to reject the language of thought hypothesis".[72]

The Churchlands' version of the type-identity theory comes from the attempt to use the resources of neoconnectionist cognitive modeling to develop a more biologically respectable form of computational functionalism. That is, they view the artificial neural networks as neurally inspired computational systems, and hence endorse the functionalist idea that the explanation of a cognitive process disregards the fact that its medium is made of nervous tissue: "Neuronal details are no more essential to connectionist conceptions of cognition than vacuum-tube or transistor details are essential to the classical conception of cognition embodied in orthodox AI, Fodorean psychology, and [folk psychology] itself".[73]

What the Churchlands blame on classical computational functionalism (aka RCTM) is that it failed to distinguish the level of cerebral *matter* from the level of cerebral *architecture*. A functionalism that aspires after biological plausibility needs to view our knowledge of the functional structure of brain as a source of constraints on the computational modeling. From this point of view, the strengths of artificial neural networks (capacities of learning and self-organization, flexibility, robustness in the presence of perturbations, capacity of dealing with such low-level tasks as the processing of sensory inputs and motor outputs) depend on just those structural