Studies in Brain and Mind 7

Carlos Muñoz-Suárez
Felipe De Brigard *Editors*

Content and Consciousness Revisited

With Replies by Daniel Dennett

Springer

Studies in Brain and Mind

Volume 7

More information about this series at http://www.springer.com/series/6540

Carlos Muñoz-Suárez • Felipe De Brigard
Editors

Content and Consciousness Revisited

With Replies by Daniel Dennett

Springer

Editors
Carlos Muñoz-Suárez
Departament de Lògica, Història i Filosofia
Universitat de Barcelona
Barcelona, Spain

Felipe De Brigard
Center for Cognitive Neuroscience
Department of Philosophy
Duke University
Durham, North Carolina, USA

The image included in the Chapter 1, according to the U.S. Copyright Policies, is in the Public Domain due to copyright expiration, because its first publication occurred prior to January 1, 1923. In this case, in 1892.

Studies in Brain and Mind
ISBN 978-3-319-17373-3 ISBN 978-3-319-17374-0 (eBook)
DOI 10.1007/978-3-319-17374-0

Library of Congress Control Number: 2015940543

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Foreword: Writing *Content and Consciousness*

Oxford in the mid-1960s dominated Anglophone philosophy as never before (and never since), and there were dozens of Americans, Canadians, Australasians, and South Africans (whites, of course, back then) eager to become certified practitioners of the then fashionable ordinary language philosophy. I was as enthusiastic as any, with Ryle's *Concept of Mind* and Austin and Wittgenstein as my beacons, but when I talked with my fellow graduate students, I discovered a disturbing complacency and lack of intellectual curiosity infecting their approaches. I remember in particular a meeting in my first term of the Ockham Society, a graduate discussion group. In the midst of a discussion of Anscombe's *Intention*, as I recall, the issue came up of what to say about one's attempts to raise one's arm when it had gone "asleep" from lying on it. At the time I knew nothing about the nervous system, but it seemed obvious to me that something must be going on in one's brain that somehow amounted to trying to raise one's arm, and it might be illuminating to learn what science knew about this. My suggestion was met with incredulous stares. What on earth did science have to teach philosophy? This was a philosophical puzzle about "what we would say," not a scientific puzzle about nerves and the like. This, it seemed to me, was as weirdly narrow an approach as setting out to learn all about horses by seeing what everyday folk had to say whenever they used the word "horse." It might help, mightn't it, to examine a few horses? My fellow philosophers of mind in Oxford were untroubled by their ignorance of brains and psychology, and I began to define my project as figuring out as a philosopher how brains could be, or support, or explain, or cause… minds.

I asked a friend studying medicine at Oxford what brains were made of and vividly remember him drawing simplified diagrams of neurons, dendrites, and axons—all new terms to me. It immediately occurred to me that a neuron, with multiple inputs and a modifiable branching output, would be just the thing to compose into networks that could learn by a sort of evolutionary process. Many others have had the same idea, of course, before and since. Once you get your head around it, you see that this really is the way—probably, in the end, the only way—to eliminate the middleman, the all-too-knowing librarian or clerk or homunculus who manipulates

the ideas or mental representations, sorting them by content. With this insight driving me, I could begin to see how to concoct something of a "centralist" theory of intentionality. (This largely unexamined alternative was suggested by Charles Taylor in his pioneering book, *The Explanation of Behaviour*.) The result would be what would later be called a functionalist, and then teleofunctionalist, theory of content in which Brentano and Husserl and Quine could all be put together, but at the subpersonal level. (The personal/subpersonal distinction was my own innovation, driven by my attempts to figure out what on earth Ryle was doing and how he could get away with it.) In order to do this right, I needed to learn about the brain, so I spent probably five times as much energy educating myself in Oxford's Radcliffe Science Library as I did reading philosophy articles and books.

I went to Ryle, my supervisor, with my project, and to my delight and surprise he recommended that I drop the arduous B.Phil. program with its brutal examinations and switch to the D.Phil., a thesis-only degree. I was off and running, but the days of inspiration were balanced by weeks and months of confusion, desperation, uncertainty. A tantalizing source of alternating inspiration and frustration was Hilary Putnam, whose *Minds and Machines* (1960) I had found positively earthshaking. I set to work feverishly to build on it in my own work, only to receive, from my mole back at Harvard, an advance copy of Putnam's second paper on the topic, "Robots: Machines or Artificially Created Life?" (not published until 1964), which scooped my own efforts and then some. No sooner had I recovered and started building my own edifice on Putnam paper number two than I was spirited a copy of Putnam paper number three, "The Mental Life of Some Machines" (eventually published in 1967), and found myself left behind yet again. So it went. I think I understood Putnam's papers almost as well as he did, which was not quite well enough to see farther than he could see what step to take next. Besides, I was trying to put a rather different slant on the whole topic, and it was not at all clear to me that, or how, I could make it work. Whenever I got totally stumped, I would go for a long, depressed walk in the glorious Parks along the river Cherwell. Marvelous to say, after a few hours of tramping back and forth with my umbrella muttering to myself and wondering if I should go back to sculpture (my alternative career path), a breakthrough would strike me, and I'd dash happily back to our flat and my trusty Olivetti for another whack at it. This was such a reliable source of breakthroughs that it became a dangerous crutch; when the going got tough, I'd just pick up my umbrella and head out to the Parks, counting on salvation before suppertime.

Ryle himself was the other pillar of support that I needed. In many regards, he ruled Oxford philosophy at the time, as editor of *Mind* and informal clearinghouse for jobs throughout the Anglophone world, but at the same time he stood somewhat outside the cliques and coteries, the hotbeds of philosophical fashion. He disliked and disapproved of the reigning Oxford fashion of clever, supercilious philosophical one-upmanship and disrupted it when he could. He never "fought back." I tried to provoke him, in fact, with some elaborately prepared and heavily armed criticisms of his own ideas, but he would genially agree with all my good points as if I were talking about somebody else and get us thinking of what repairs and improvements we could make together of what remained. It was disorienting, and my opin-

ion of him then—often expressed, I am sad to say, to my fellow graduate students—was that while he was wonderful at cheering me up and encouraging me to stay the course, I hadn't learned any philosophy from him.

I finished a presentable draft of my dissertation in the minimum time (six terms or 2 years) and submitted it, with scant expectation that it would be accepted on first go. On the eve of submitting it, I came across an early draft of it and compared the final product with its ancestor. To my astonishment, I could see Ryle's influence on every page. How had he done it? Osmosis? Hypnotism? This gave me an early appreciation of the power of indirect methods in philosophy. You seldom talk anybody out of a position by arguing directly with their premises and inferences. Sometimes it is more effective to nudge them sideways with images, examples, and helpful formulations that stick to their habits of thought.

My examiners were A. J. Ayer and—an unprecedented alien presence at a philosophy "viva" occasioned by my insistence on packing my thesis with speculations on brain science—the great neuroanatomist, J. Z. Young from London. He, too, had been struck by the idea of learning as evolution in the brain and was writing a book on it, so we were kindred spirits on that topic, if not on the philosophy, which he found intriguing but impenetrable. Ayer was reserved. I feared he had not read much of the thesis, but I later found out he was simply made uncomfortable by his friend Young's too-enthusiastic forays into philosophy, and he found silence more useful than intervention. I waited in agony for more than a week before I learned, via a cheery postcard from Ryle, that the examiners had voted me the degree.

I returned to the United States, to UC Irvine, my first teaching job, age 23. Now it was time to turn my dissertation into articles and a book. I revised the first chapter and sent it out as a journal article. There were a dozen submissions and a dozen rejections, with many revisions in between. Then Wilfrid Sellars, editor of *Philosophical Topics*, wrote me a nice letter saying that he was intrigued by the draft I had sent; once I clarified a few foggy points, he thought it would be fine. I sent him a clarified version within the week, and he wrote back to say that now that it was clear what I was doing, he thought it was not a publishable paper! A few more rejections and I gave up on that chapter and started several other projects, with no greater success. Perhaps I wasn't going to make it as a philosopher after all.

One day Julian Feldman, an artificial intelligence researcher at UCI, came storming into my office with a copy of Hubert Dreyfus's notorious RAND memo, "Alchemy and Artificial Intelligence." What did I make of it? I read it and said I disagreed quite fundamentally with it. "Write up your rebuttal, please, and get it published!" Why not? I wrote "Machine Traces and Protocol Statements" and promptly published it in *Behavioral Science* (1968, my first publication), and my career as philosopher-laureate of AI had begun. I'd already been attracted to the field by Alan Ross Anderson's pioneering anthology, *Minds and Machines*, and found at Irvine a small group of AI researchers who invited me to join them. Allen Newell came through town to give some talks and struck up a lively conversation with me, and I was hooked. Other colleagues at Irvine, in particular the psychobiologist James McGaugh, struck by my knowledge of, and interest in, theories of learning in neural systems, also took a vigorous interest in further educating me and getting me thinking about their work and its problems.

I decided I had to concentrate on turning the dissertation into a book, and I think these nonphilosophers contributed the most to the improvements, clarifications, and enlargements that distinguish *Content and Consciousness* from the naive stumblings in my D.Phil. dissertation.

In the summer of 1967, I sent the new manuscript to the famous Routledge & Kegan Paul series, the International Library of Philosophy and Scientific Method. This series of books, with their red covers and yellow dust jackets, included most of my favorite books in philosophy: Wittgenstein's *Tractatus*, Smart's *Philosophy and Scientific Realism*, and Sellars' *Science, Perception and Reality*, for instance. A year passed without a word from the new editor, Ted Honderich, who had taken over from A. J. Ayer. I didn't dare upset the applecart by complaining about the unresponsiveness. Finally, when I knew I was off to Oxford for a quarter sabbatical in the fall of 1968, I wrote a timid inquiry to Honderich, who discovered that the manuscript had been mislaid by the referee to whom he had sent it. He retrieved it, read it himself, and forthwith accepted it, pending revisions which I hastened to complete that autumn in Oxford. I was in heaven. But still, I couldn't talk about it to other philosophers. My problem was that my way of approaching the then standard issues in the philosophy of mind was too eccentric, too novel, to afford easy entry into a discussion. When somebody asks you what you're working on, you usually can't back them into a corner and harangue them for a couple of hours about your project, and I could imagine no more modest framing job that might bring interlocutors to where I was. After all, my attempts to publish the first chapter showed that the first ideas I needed to get across were bound to be misunderstood and had already been misunderstood in half a dozen versions by some of the best philosophers of mind in the field. So I was a very lonely and uncertain philosopher those first few years at Irvine, spending more happy hours talking AI or psychobiology than philosophy. In spite of my presumably sterling pedigree as a student of Quine and Ryle, I felt like an outsider, a dark horse candidate that one should probably not bet on. The acceptance of the manuscript by Honderich, and his further invitation to write an essay on free will ("Mechanism and Responsibility," in which I introduced the terminology of the intentional stance and intentional systems), gave me new confidence, however.[1]

When *Content and Consciousness* was published, in 1969, J. Z. Young sent me a nice note telling me to ignore the review in the *Times Literary Supplement*, which I hadn't seen until he drew my attention to it. This was my first review, and it was a stinker. Reviews in the *Times Literary Supplement* those days were all anonymous, but years later I learned that it had been written by D. W. Hamlyn, and to my dismay one of his chief criticisms was about the style, which I had thought to be refreshingly unlike other philosophy books of the day. Young's note did cheer me up, however, and soon the book got two wonderful reviews: J. J. C. Smart did a long "Critical Notice" in *Mind*, and R. L. Franklin wrote an even more positive long

[1] The preceding paragraphs are drawn, with minor revisions, from my essay "Autobiography," published in *Philosophy Now* (London, July 2008).

review in *Australasian Journal of Philosophy*. These two reviews put my book in the limelight, and soon I began to field inquiries and invitations from all over the Anglophone philosophical world. Gilbert Harman at Princeton was one of the book's first supporters, as I learned when a former student of mine who had gone on to Princeton wrote me a note telling me that my book was being featured in his course. Richard Rorty was another enthusiastic reader. Princeton, in fact, was the epicenter of interest, and I was invited to give a talk there in December of 1970, the first professional talk of my career. (I presented "Intentional Systems" to an audience that included, in addition to Harman and Rorty, Alonzo Church, Donald Davidson, David Lewis, Thomas Nagel, Max Black, and quite a few other luminaries. I was terribly nervous, but the reception was cordial and constructive. Rorty had a reception for me afterwards at his house, beginning a lifelong friendship.)

The book was chosen for an "author meets critics" session at the annual meeting of the Eastern Division of the American Philosophical Association in December of 1972 more than 3 years after it was published—the world moved more slowly then. Michael Arbib, one of the first computational neuroscientists, and Keith Gunderson, philosopher and poet, were the critics. (I think Arbib's and Gunderson's talks were never published, but my response is included as a chapter in Brainstorms.) Since my book was thus featured in a symposium, I expected the American publisher, Humanities Press, to have it prominently displayed at their table in the book exhibit room, but to my dismay they didn't have a single copy to show or sell. When I confronted the proprietor with this anomaly, his response aggravated my bad mood: "A symposium? So that's why people have been coming around all day asking for it!" He had no copies because it wasn't a new book. I later learned, moreover, that Routledge & Kegan Paul, which had an arrangement with Humanities Press to print their copies with a different front page bound in, had decided some months before the symposium that my book wasn't going anywhere and had remaindered the rest of their stock to Humanities Press for a dollar a copy. So when the book did take off and have a good sale in the United States, I got the handsome royalty of ten cents a copy. But that didn't matter to me; the book was being read and discussed in courses and seminars.

One of my favorite responses to the book came from Arthur Danto, whom I had not yet met. He sent me a nice note about how much he had enjoyed the book, and learned from it, but then he went on to draw my attention to one of the embarrassing errors in it. I had misexplained Quine's famous example "Giorgione was so-called because of his size." I had supposed that Giorgione meant Little George, not Big George. Danto enclosed a copy of the letter he had just sent to Quine, informing him that in any case Quine was wrong, too! According to Vasari, Danto noted, Giorgione was so-called dalle fattezze dalla persona e dalla grandezza del animo—because of the features of his face and the greatness of his soul. "However," Danto went on, graciously, "it is not my intention to wander either into questions of physiognomy or grammar, but to report a factual error which would be minor in the case of someone who did not bear the awful responsibility of stocking philosophers with what meager facts they may claim." I immediately wrote to Quine, apologizing for butch-

ering his example, but taking some pleasure in Danto's discovery that Quine himself had perpetrated a falsehood. Quine immediately wrote back a friendly letter, enclosing a copy of his reply to Danto:

Dear Danto,

> I much appreciated your generous and amusing letter of August 17. I find a perceptible gain in taking on a new fact and getting rid of a dud, whatever the chagrin over being caught out. But the present case leaves me in doubt. My dictionary gives "fattezze" as "bodily proportions," among other things… Vasari softens the blow by ringing the animo in too, but I would set that down to the animo of Vasari.

Over the years Quine delighted in finding embarrassing factual errors in my books, and it became a running joke. I never caught him out in a factual error, but when I sent him a list of the factual errors in *Darwin's Dangerous Idea* (the book dedicated to him) that he had missed, we had a good laugh over it.

It seems to me that for all its flaws, *Content and Consciousness* had enough things right to make it an excellent platform on which to build further philosophical work. Or better, it has been, for me, a sort of philosophical kitchen, stocked with almost all the utensils and containers, all the ingredients and methods, from which I have concocted the rest of my work. And over the years I have enjoyed watching other philosophers gravitating inexorably towards versions of the views I first spelled out there. What seemed outrageous and even incomprehensible to many of my colleagues 40 years ago makes much more sense today.

Medford, MA, USA Daniel Dennett

Acknowledgments

The Editors: We thank Gualtiero Piccinini, the editor of this excellent book series, for his constant patience and invaluable help. We want to thank also the authors of the articles included in this volume, as well as those who helped review their manuscripts.

Carlos Muñoz-Suárez: I'm grateful to Dan Dennett for his support and willingness to discuss and update his ideas. It has been a pleasure to work with Felipe De Brigard. I thank him for his remarkable patience and outstanding collaboration. I also thank Pete Mandik and Don Ross for their support from the earliest stage of this project. I'm specially grateful to Aura Marina Suárez for her unconditional love and support. To my family and friends, thanks for sailing with me across the seas of love and knowledge. Finally, I thank the PERSP Project (Consolider-Ingenio project CSD2009-00056) for funding my life in Barcelona.

Felipe De Brigard: First, I would like to thank Dan Dennett for being a source of philosophical inspiration and for his constant support throughout the years. Thanks also to Carlos Muñoz-Suárez, my coeditor, for getting this project going. Finally, I would like to thank my wife, Anne, my son, David, the one the way, and our dog, Jeffers, for their unending love. You guys are awesome.

Contents

Contributors

Felipe De Brigard Center for Cognitive Neuroscience, Department of Philosophy, Duke University, Durham, NC, USA

Daniel Dennett Tufts University, Medford, MA, USA

Richard Dub University of Geneva, Geneva, Switzerland

Keith Frankish The Open University, Milton Keynes, UK
The University of Crete, Heraklion, Greece

Ellen Fridland King's College London, London, UK

Pete Mandik William Paterson University of New Jersey, Wayne, NJ, USA

John Michael Central European University, Budapest, Hungary

Carlos Muñoz-Suárez Departament de Lògica, Història i Filosofia, Universitat de Barcelona, Barcelona, Spain

Don Ross University of Waikato, Hamilton, New Zealand
University of Cape Town, Cape Town, South Africa
Georgia State University, Atlanta, GA, USA

Martin Roth Drake University, Des Moines, IA, USA

Sam Wilkinson Durham University, Durham, UK

About the Contributors

Carlos Muñoz-Suárez is a doctoral researcher at the Perspectival Thoughts and Facts Project and PhD member at the Logos Group in Analytic Philosophy at the University of Barcelona. He earned a BA in philosophy, a BA in psychology, and an MPhil in philosophy of mind and cognitive science. He has been a lecturer in philosophy and psychology at some universities in Colombia. His research interests include visual consciousness, visual information, visually guided behavior, and the metaphysics of visible things. He is also interested in the psychology and philosophy of intuitions.

Don Ross is a professor of economics and dean of commerce at the University of Cape Town and a program director for methodology at the Center for Economic Analysis of Risk at Georgia State University. He has recently been appointed dean of the Waikato Management School and professor of economics at the University of Waikato, New Zealand, a post he will take up in 2015. His areas of recent research include economic methodology, experimental economics of risk and time preferences in vulnerable populations, strategic foundations of human sociality, and scientific metaphysics. His many publications include *Economic Theory and Cognitive Science: Microexplanation* (2005), *Every Thing Must Go: Metaphysics Naturalized* (with J. Ladyman) (2007), *Midbrain Mutiny: The Picoeconomics and Neuroeconomics of Disordered Gambling* (with C. Sharp, R. Vuchinich and D. Spurrett) (2008), and *Philosophy of Economics* (2014).

Ellen Fridland is a lecturer in philosophy at King's College London. Before joining King's, she was a visiting fellow with Dan Dennett at the Center for Cognitive Studies at Tufts University. She works primarily on skill.

Felipe De Brigard is an assistant professor at Duke University in the Department of Philosophy, the Center for Cognitive Neuroscience, and the Duke Institute for Brain Sciences (DIBS), where he directs the Imagination and Modal Cognition Lab. He earned a BA from the National University of Colombia, an MA from Tufts University, and a PhD from the University of North Carolina, Chapel Hill. He spent 2 years as

postdoctoral fellow at the Cognitive Neuroscience of Memory Lab at Harvard University. His research centers on the interaction between memory and imagination and the relationship between attention, consciousness, and recollection.

John Michael is a Marie Curie research fellow at the Somby Lab within the Department of Cognitive Science of the Central European University. After completing his PhD in philosophy at the University of Vienna, he worked as a postdoc in cognitive science at Aarhus University, Copenhagen University, and Tufts University. His background is in philosophy of mind, cognitive science, and philosophy of science, and his main interests are in social cognition research.

Keith Frankish is a visiting senior research fellow at the Open University UK and an adjunct professor with the Brain and Mind Program in Neurosciences at the University of Crete, Greece. He is the author of *Mind and Supermind* (2004) and *Consciousness* (2005), as well as numerous articles and book chapters. He is coeditor of *In Two Minds: Dual Processes and Beyond* (with Jonathan St. B. T. Evans, 2009), *New Waves in Philosophy of Action* (with Jesús H. Aguilar and Andrei A. Buckareff, 2010), *The Cambridge Handbook of Cognitive Science* (with William M. Ramsey, 2012), and *The Cambridge Handbook of Artificial Intelligence* (with William M. Ramsey, 2014). His research interests include dual-process theories of reasoning, the psychology of belief, and the nature of phenomenal consciousness.

Martin Roth is an associate professor of philosophy at Drake University. His research interests include mental representation, functional explanation, and knowledge-how, and his work has appeared in *Synthese*, *Philosophical Studies*, *Philosophical Psychology*, *Mind and Language*, and other journals and collections.

Pete Mandik is a professor of philosophy at William Paterson University in New Jersey, USA. He specializes in philosophy of mind and philosophy of science. He is cohost of the podcast *SpaceTimeMind*. He has published *This Is Philosophy of Mind: An Introduction* (author, Wiley-Blackwell, 2013), *Key Terms in Philosophy of Mind* (author, Continuum, 2010), *Cognitive Science: An Introduction to the Mind and Brain* (coauthor, Routledge, 2006), and *Philosophy and the Neurosciences: A Reader* (coeditor, Blackwell, 2001).

Richard Dub is a postdoctoral research fellow at the Swiss Center for Affective Sciences in Geneva, Switzerland. He received his PhD in philosophy from Rutgers University in 2013. His research concerns psychotic delusions and pathological beliefs, an area of philosophy that he was inspired to study after attending a course of Daniel Dennett's in the philosophy of psychiatry.

Sam Wilkinson is a postdoctoral research fellow at Durham University. He received his PhD from the University of Edinburgh for a thesis entitled "Monothematic Delusions and the Nature of Belief." Before that, he did his BA at St. Catherine's College, Oxford University, and a master's at the Jean Nicod Institute in Paris. He currently works within a Wellcome Trust-funded project that examines auditory verbal hallucinations in both pathological and non-pathological contexts.

Chapter 1
Introduction: Bringing Together Mind, Behavior, and Evolution

Carlos Muñoz-Suárez

> *The very same tree that Tommy could not climb last year is climbed by him this year because his legs and arms are longer. So, not indeed the tree, but his task has changed. Thus too the thinker, the converser or the fencer is himself, in some measure, a once-only factor in his own once-only situations. It would be absurd to command him 'Think again exactly what you thought last time'; 'Repeat without any change at all your experiment of last time'. The command itself would be a fresh influence. To obey it would be disobey it.*
>
> -RYLE, G. 1969: 130-

Abstract In Sect. 1.1 I discuss the main concepts and hypotheses introduced in *Content and Consciousness*. In Sect. 1.2 I sketch the context of interdisciplinary research surrounding *Content and Consciousness's* birth. Finally, in Sect. 1.3, I introduce the chapters of this volume.

Content and Consciousness (hereafter, *C&C*) is widely recognized as a pioneering work that provided a framework for an account of mind and behavior developed through the unification of scientific findings. *C&C* initiated several contemporary research trends in philosophy and science. How this happened is a fascinating story. The present volume is devoted to revisiting the hypotheses, concepts and distinctions introduced in *C&C* from an updated interdisciplinary perspective.

C&C sowed the seeds of philosophical gardens, where pioneering ideas and ideals flourished. Those gardens outgrew their boundaries, spreading their seeds on neighboring scientific fields. Philosophical and scientific vegetation became entangled and hybridized. Almost five decades later, what once looked like an abyss between those lands is now a barely discernible crack.

C. Muñoz-Suárez (✉)
Departament de Lògica, Història i Filosofia, Universitat de Barcelona, Barcelona, Spain
e-mail: carlosmariomunozsuarez@gmail.com

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*, Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_1

This introduction is divided into three sections. In Sect. 1.1 I discuss in some detail the main ideas of *C&C*. In Sect. 1.2 I briefly sketch the context of interdisciplinary research surrounding *C&C*'s birth. Finally, in Sect. 1.3, I summarize the chapters of this volume.

1.1 Seven Seeds

1.1.1 A Forerunner of the Intentional Stance

In general, *C&C*'s analysis of intentionality concerns what assumptions should be rejected in order to dissolve the classic mind-body problem. The main goal of *C&C* was to provide "a scientific explanation of the differences and similarities in what is the case in virtue of which different mental language sentences are true and false" (18–19).[1] With that goal in mind, Dennett developed a scientifically-inspired philosophical theory as well as a philosophically-framed scientific account of mind and behavior.

Inspired by Chisholm (1957) and Quine (1960), Dennett introduced an analysis of mental vocabulary designed to discredit widely held metaphysical assumptions deriving from Descartes' epistemology and Brentano's psychology.[2] Notably (and very much in the spirit of Wittgenstein, Ryle and Quine), Dennett did not adopt that linguistic approach merely to engage in traditional conceptual analysis, but rather in order to avoid spurious philosophical puzzles and break through the barriers that orthodox philosophers had built around the empirical sciences, thereby opening a path to scientific progress in philosophical inquiry.

According to *C&C*, philosophers should investigate to see which terms of mental vocabulary (if not all) are non-referential,[3] since "[w]hat we start with […] are sentences containing the mental entity words to be examined […] The broader question […] is whether or not these sentences, accepted either as wholes or as analysed, can be *correlated in an explanatory way* with sentences solely from the referential domain of physical sciences" (16–17). Dennett thus aimed to promote ontological neutrality (15) in the analysis of mental sentences. Success here would absolve, on the one hand, scientists (mainly neuroscientists) "from the responsibility of discovering physical events, states or processes which deserve to be called thoughts, ideas,

[1] Most references to *C&C* will be solely indicated with the page numbers from the 1986 edition (Dennett 1986). When context requires it the page numbers will appear following '*C&C*:'.

[2] Accordingly Dennett claimed: "The first step in finding solutions to the problems of mind is to set aside ontological predilections and consider instead the relation between the mode of discourse in which we speak of persons and the mode of discourse in which we speak of bodies and other physical objects" (189).

[3] "Non-referential words and phrases are then those which are highly dependent on certain restricted contexts, in particular cannot appear properly in identity contexts and concomitantly have no ontic force or significance. That is, their occurrence embedded in an asserted sentence never commits the asserter to the existence of any entities presumed denoted or named or referred to by the term" (14).

mental images and so forth" (19) and, on the other hand, physicalist philosophers from justifying their identifications of physical and mental entities. This approach challenged the view that there are physical correlates for each meaningful category of our mental vocabulary – a view motivating what was often called in traditional neuropsychology "naive localizationism", or "the phrenology of the 20th century",[4] which was explicitly promoted by some traditional identity theorists in philosophy (Place 1956; Feigl 1958).

Although Dennett favored relaxing the ontological assumptions associated with mental vocabulary, he didn't dispute its explanatory role. Indeed, he emphasized that no science of behavior can get along without mental vocabulary (34), since, in principle, "no sentence or sentences can be found which adequately reproduce the information of an Intentional sentence" (30). This idea was clearly a forerunner of his famed intentional stance, to be discussed below.

According to *C&C*, intentional sentences have value for making sense of (or rationalizing) complex system's inner and outer responses. Nonetheless, the explanation of mind and behavior cannot be developed purely on the basis of intentional considerations. In fact, a fundamental motivation for the forerunner of the intentional stance found in *C&C* was to overcome the explanatory incompleteness of a radical (stimulus-response) behavioral science, as well as that of a purely intentional science. Dennett thus accounted for the role of intentional sentences in the development of a complete science of mind and behavior, showing that both radical behaviorism and pure intentionalism were misguided and misleading.

According to radical behaviorists such as Skinner (1938, 1957), the contingent connections between environmental conditions (captured by the antecedents of stimulus-response conditional statements) and behavior (captured by their consequents) constituted the *explanandum* of the science of behavior.[5] Radical behaviorists thought that the environmental appropriateness of behavior was all that a scientific "psychological" theory should explain. By contrast, according to *C&C*, a science of mind developed solely on the basis of peripheralist stimulus-response statements (43) would not be *complete*, since its statements would exclude the information carried by intentional descriptions and hence ignore their putative role in rationalizing behavior and systems' activities in general. In short, Dennett argued for the explanatory significance of intentional descriptions in a complete science of mind and behavior.

In criticism of pure intentionalism, Dennett remarked that purely intentional explanations[6] conflict with the explanatory structure of the causal statements commonly found in empirical sciences. In particular, according to pure intentionalists, in the statements of a complete science of mind, the connection between anteced-

[4] For a detailed discussion of this view, see Uttal 2003.

[5] In this vein, Dennett claimed: "[f]or the super-abstemious behaviorist who will not permit himself to speak even of intelligence (that being too 'mentalistic' for him) we can say, with Hull, that a primary task of psychology 'is to understand… why… behavior… is so generally adaptive, i.e., successful in the sense of reducing needs and facilitating survival…'" (Dennett 1981b: 72).

[6] For instance: 'I believe in ghosts *since* I have seen them', 'she raised her arm *because* she wanted to ask a question' and 'his belief in the bogeyman *prompted* her to look at inside the closet.'

ents (for instance, 'She intends to open the door') and consequents (for instance, 'She opens the door') appears to be a priori, thus playing no role in a Humean causal explanation (37–38) and misrepresenting the empirical character of standard explanations in the empirical sciences.[7] Consequently, a science exclusively developed on the basis of intentional statements shouldn't be viewed as empirically supported. Hence *C&C* articulated a view on which intentional statements earned their legitimacy *in service* of an empirically supported science.

According to radical behaviorism, scientific psychological explanations should exclude intentional statements. Meanwhile, according to pure intentionalism, intentional statements are essential for accounting for mental phenomena. The analysis of mental vocabulary provided by these views, as in other proposed answers to the mind-body problem, is framed by certain ontological predilections. By contrast, *C&C* vindicates a stance of ontological neutrality in its hypothesis that many (perhaps all) terms of the mental vocabulary are plausibly non-referential.

The forerunner of the intentional stance was introduced as a revision of Brentano's view of intentionality. According to Brentano:

> Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on […] This intentional in-existence is characteristic exclusively of mental phenomena. *No physical phenomenon exhibits anything like it*. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves. (Brentano 2009: 68. Italics mine)

The first dialectical move in *C&C* was to "relax" the ontological predilections (189) framing the use of intentional terms by drawing an analogy with certain non-referential terms – e.g., 'voice' and 'sake' (8 and ff.).[8] In this way, intentional terms (like 'pain', 'desire', and so on), *pace* Brentano, need not be taken to refer to non-linguistic phenomena bearing the peculiar feature of intentionality.

C&C's revision of Brentano's view avoided any proclivity to postulate queer kinds of entities, like Brentanian in-existent entities, Meinongnian objects, sense data, and Cartesian ghosts. Trying to bridge the gap between such queer non-physical entities and the entities postulated by the empirical sciences was precisely what pushed philosophers into a blind alley in which they had no option other than endorsing dualism (interactionist or epiphenomenalist), monist materialism or monist idealism.

In contrast with Brentano's view, in *C&C*:

> Intentionality is not a mark that divides phenomena from phenomena, but sentences from sentences […] Brentano's thesis might be altered to read '*All mental phenomena are directed by (or simply: related to) unique descriptions or whole propositions which usually, but not always, have reference to real objects in the world*.' Thus Brentano's thesis becomes

[7] Although "[i]ntentional explanations explain a bit of behavior, an action, or a stretch of inaction, by making it reasonable in the light of certain beliefs, intentions, desires ascribed to the agent" (Dennett 1981c: 236).

[8] An expression is 'non-referential' not due to its having non-existent referents, like some philosophers claim about 'unicorn'. By contrast, an expression is non-referential in case of being semantically embedded in sentences it appears. See: fn. #3, and *C&C*: 13, fn. #1.

> […] simply that mental phenomena differ from physical phenomena in having a content, or relating to meaning, *in the sense that their identity as individual phenomena is a matter of the unique descriptions or propositions to which they are related* […] *Raising the subject level of discussion back up from phenomena to talk about phenomena, from things to sentences*, the point is this: Intentional sentences are *intensional* (non-extensional) sentences. (27, 29. Italics mine)

According to this view, systems' internal phenomena have content insofar as their users and observers use intentional descriptions for making sense of their activities. Dennett argued that ascribing content does not consist in locating intentional phenomena within systems, nor in identifying such phenomena with internal physical states, events or processes. So ascribing content involves neither the reduction nor the reification of intentional entities. In this way, accounting for the relation between intentional descriptions and extensional phenomena doesn't consist in relating intentional phenomena to extensional phenomena, but in understanding the relations between intentional sentences and extensional phenomena.

The seed of the intentional stance was sowed in the following terms:

> A computer can only be said to be believing, remembering, pursuing goals, etc., relative to the particular interpretation put on its motions by people, who thus impose the Intentionality of their own way of life on the computer. That is, no electrical state or event in a computer has any intrinsic significance, but *only the significance gifted it by the builders or programmers who link the state or event with input and output*. Even the production of ink marks on the output paper has no significance except what is given it by the programmers. Thus computers, if they are Intentional [systems], are only Intentional in virtue of the Intentionality of their creators. (40–41. Italics mine)

Some years later on, Dennett introduced the label ('intentional stance') as follows[9]:

> Prediction from the intentional stance assumes rationality in the system, but not necessarily perfect rationality […] Whenever one can successfully adopt the intentional stance toward an object, I call that object an *intentional system*. The success of the stance is of course a matter settled pragmatically, without reference to whether the object *really* has beliefs, intentions, and so forth; so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably *are* intentional systems, for they are systems whose behavior can be predicted, and most efficiently predicted, by adopting the intentional stance toward them […] This tolerant assumption of rationality is the hallmark of the intentional stance with regard to people as well as computers. (Dennett 1981c: 238)[10]

Dennett thus argued that content ascriptions are "a heuristic overlay on the extensional theory rather than intervening variables of the theory". (80)[11]

---

[9] Dennett developed this view in more detail after *C&C* (Dennett 1981d, e, f) and particularly in *The Intentional Stance* (1987).

[10] For a penetrating analysis of this rationality requirement, see: Chap. 5 of this volume. As Dennett says in the "Preface to the Second Edition" of *C&C*, in *C&C* the term 'intentional system' appears in several occasions, "but not with the precise sense [Dennett] later developed" in (1981d).

[11] According to the notion of 'intentional stance,' some artificial devices and organisms would count as intentional systems only if they make intelligent use of information and their activities are intentionally rationalized. "Intentional explanations have the actions of persons as their primary domain, but there are times when we find intentional explanations (and predictions based on them) not only useful but indispensable for accounting for the behavior of complex machines" (Dennett 1981c: 236–237). See: Dennett (1981b: 80 and ff).

Rationalizing activities with intentional notions (i.e., by adopting the intentional stance) is not something that our brains can do alone, but rather something that whole observers – persons – do. The main reason for this is that adopting the intentional stance requires knowledge of environmental conditions, informational processing events and behavioral responses, to which brains alone do not have access. In particular, each brain is "'blind' to the external condition[s] producing its input[s]" (48), only having access to the information within the system in which it takes part. Therefore, the adoption of the intentional stance as a heuristic strategy is available only to certain observers or users – *ex hypothesi*, those able to deploy intentional notions and access certain knowledge.

Adopting the intentional stance toward the activities of a system is not enough to specify the correctness conditions of content ascriptions. The task of specifying such correctness conditions requires relating intentional descriptions to a well-established scientific framework that provides an extensional description of systems' structures, functions and origins.

As a matter of principle, systems' status as intentional cannot be bestowed by a supremely intelligent creature. The postulation of such a supreme intelligence begs the question regarding the origins of intelligence in the natural order: it would explain intelligence by presupposing intelligence (Dennett 1981b: 72). Hence, by an inference to the best explanation, "we must find something else to endow [systems'] internal states with content [… W]e can look to the theory of evolution by natural selection" (41). Evolutionary naturalism "enters the scene" as a foundational view that can frame a scientifically respectable theory of the correctness conditions of content ascriptions. Why should we favor evolutionary naturalism?

1.1.2 A Motivation for Evolutionary Naturalism

Dennett's naturalism can be seen in how he specifies correctness criteria for mental language in scientific language. However, his naturalism should not be taken to support a reductive program: it is not aimed at analyzing intentional notions in purely extensional vocabulary. His naturalism rather consists in *legitimizing* mental vocabulary within a scientific framework.[12]

The thread of evolutionary naturalism runs through the whole of *C&C*, though the book doesn't contain a direct defense of Darwin's theory of evolution and its place in the science of mind.[13] Rather the theory of evolution is used here as the best available meta-theoretical framework – "from the standpoint of the extensional, physical theory" (80) – for specifying the correctness conditions of content ascriptions.

[12] Or, as Dennett later claimed, in a regimentation of mentalistic notions (Dennett 1981a: xix). For an interpretation of the relation between philosophy and scientific theories in *C&C*, see the Chap. 2 of this volume.

[13] Such a motivation could be found in Dennett's (1981b). His most detailed discussion about evolutionary naturalism is in his *Darwin's Dangerous Idea* (1995).

Obtaining an optimal rationalization of systems through the intentional stance depends on ascribing to them *discriminations* – for instance, of features of their environment – on the basis of empirical criteria regarding their functions and structures, and specified by looking to their outer and inner evolutionary[14] and developmental histories. Thus evolutionary naturalism enters the scene.

Systems' environmental discriminations and their associated behavioral responses have *survival value*. Dennett illustrated this idea as follows:

> Suppose that there were three different strains of a certain primitive organism in which a certain stimulation or contact caused different 'behaviour'. In strain A the stimulation happened to cause the organism to contract or back off; in strain B the only behaviour caused by the electrical activity in it was a slight shiver or wriggling; in strain C the stimulation caused the organism to move towards or tend to surround or engulf the point of contact causing the stimulation. Now if the stimulation in question happened to be caused more often than not by something injurious or fatal to the organism, strain A would survive, strain B would tend to die off and strain C would be quickly exterminated (other conditions being equal). But if the stimulus happened to be caused more often than not by something beneficial to the organism, such as food, the fates of A and C would be reversed. Then, although all three responses to the stimulation are blind, the response that *happens* to be appropriate is endorsed through the survival of the species that has this response built in. (49)

According to evolutionary naturalism, specifying criteria for the correctness of content ascriptions requires seeing whether those ascriptions optimally reflect the discriminations that the relevant systems actually make – which further requires knowing whether those systems have functional structures enabling an intelligent use of information. Natural selection is then used to explain the appropriateness of discriminatory activities.

The main lesson of evolutionary naturalism is that behavioral appropriateness derives from evolutionary success. "Any afferent-efferent connection that was regularly appropriate would have survival value, the likelihood of survival depending on how regular the beneficial environmental results of the response motion are" (50). So natural selection "guarantees, over the long run, the environmental appropriateness of what it produces" (41), and the "species that survive are the species that happen to have output of efferent impulses connected to the afferent or input impulses in ways that help them to survive [… T]he organisms that survive will be those that happen to react differently to different stimuli – to discriminate" (49).

According to *C&C*, the search for a scientific framework for evaluating content ascriptions unavoidably leads to a journey into the evolutionary and developmental histories of systems. However, we shouldn't merely investigate the evolution of *species* of systems, but also the "intra-cerebral evolutionary processes" (60) by which their brains' functional structures[15] were selected to enable their environmentally embedded activities. In the next section I summarize the main tenets of this hypothesis.

---

[14] See: fn # 17.

[15] According to *C&C*, a functional structure is "any bit of matter (e.g., wiring, plumbing, ropes and pulleys) that can be counted on – because of the laws of nature – to operate in a certain way when operated upon in a certain way […] A functional structure can break down – not by breaking laws of nature but by obeying them – or operate normally" (48). This notion applies to the behavioral

1.1.3 Learning as Intra-Cerebral Evolution

According to *C&C*, brains rely on their highly dynamic capacity to react "differentially to stimuli in appropriate response to the environmental conditions they herald" (47). A brain lacking that capacity cannot serve its system. According to Dennett, neither bodily reactions nor internal states (e.g., electrical patterns) have intrinsic significance. Thus we may ask: how is the brain to discriminate stimuli in accordance with their environmental significance? How do brains stabilize their functional structures for this purpose? Part of the answer lies in understanding learning as intra-cerebral evolution.

This seed was firmly planted by *C&C*.[16] Rather than understanding evolution just as a mechanism by which species appear in and disappear from their *outer* environments, it should be more broadly understood as a mechanism for *informational selection* – of species in outer environments as well as of functional structures in *inner* environments.[17] The hypothesis of intra-cerebral evolution entered the neuroscientific research agenda after *C&C*[18] (see: Changeux et al. 1973; Changeux and Danchin 1976; Edelman 1987; Changeux and Dehaene 1989).

The hypothesis of learning as intra-cerebral evolution derived from an account of how the discriminations made by various systems improve their fitness, as well as of the *appropriate structures* for realizing such discriminations. According to *C&C*,

> [s]ince environmental *significance*, even in the attenuated sense in which retinal impulse streams *signify* certain retinal conditions, is not an intrinsic physical characteristic, the brain, as a physical organ, cannot sort by significance by employing any physical tests. The only other explanation that would be acceptable to the physical sciences is that the brain's capacity to discriminate appropriately is based on chance. That is, a particular pathway through the brain might just happen – entirely fortuitously – to link an afferent (input) event

---

control system of natural as well as of artificial systems, like computer programs. Here, as in several places in *C&C*, "the strength of the analogy between human behaviour and computer behaviour is […] a critical point" (45). Moreover, functional structures are compound afferent-efferent informational patterns that are realized by a multitude of "switching elements" (e.g., neurons) which have the capacity to propagate and stabilize informational and physical pathways (52, 54).

[16] It is worth-mentioning that Hebb (1949) introduced, from a biopsychological standpoint and a neurological talk, associated hypotheses about the relations between structural-functional brain changes and learning processes. Ross (see: fn. #4 in Chap. 2 of this volume) claims that Dennett knew Hebb's work when he wrote *C&C*.

[17] Dennett claims:

> […] creatures have *two* environments, the outer environment in which they live, and an "inner" environment they carry around with them […] it is environmental effects that are the measure of adaptivity and the mainspring of learning, but the environment can delegate its selective function to something in the organism (just as death had earlier delegated its selective function to pain), and if it occurs, a more intelligent, flexible, organism is the result. (Dennett 1981b: 77, 78)

[18] Forty years after *C&C*, Dennett claimed that "[o]nce you get your head around [this idea], you see that this really is the way – probably, in the end, the only way – to eliminate the middleman, the all-too-knowing librarian or clerk or homunculus who manipulates the ideas or mental representations, sorting them by content" (Dennett 2008).

> or stimulus to an efferent (output) event leading to appropriate behaviour, and if such fortuitous linkages could in some way be generated, recognized and preserved by the brain, the organism could acquire a capacity for generally appropriate behaviour. (48)

The environmental significance that input conditions have *for* a particular system is neither endowed by its users and observers, nor exclusively determined by its pre-efferent activity. Rather, "discrimination of afferents according to their significances just *is* the production of efferent effects in differential response to afferents, and hence it does not make sense to suppose that prior to the production of an efferent event or structure the brain has discriminated its afferents *as* anything at all" (74). Therefore, ascribing such discriminations (in a principled way) depends on understanding the whole interplay between afferent and efferent information-processing events in the relevant system.

Understanding how systems consume information helps us to understand how they discriminate their inputs by their environmental significance. The appropriateness of those discriminations depends on how systems' behavioral responses fulfill their needs, given the particular environmental situations in which they are embedded, i.e., depends on the intelligent use of information by the relevant systems.[19]

Dennett provided an evolutionary account of intra-systemic structures, which was the core of his hypothesis of learning as intra-cerebral evolution. He argued that "[w]hat is needed is some intra-cerebral function to take over the evolutionary role played by the exigencies of nature in species evolution; i.e., some force extinguish the inappropriate [responses]" (52). In short, as species conflict for surviving in their outer environments, individual systems' functional structures conflict for being sorted in "inner" environments (57). Such "inner" environments are fabrics of afferent-efferent patterns, enabling individual discriminatory capabilities and behavioral responses.

Dennett appealed to intra-cerebral evolution, on the one hand, to explain the grounds of learning and intelligence without presupposing intelligence. In this way, intra-cerebral evolution endows systems with flexible capacities required to make intelligent use of information, for instance, required for predicting rewards without actual behavior (*vid.*: Dennett 1981b: 79 and ff.). On the other hand, Dennett appealed to intra-cerebral evolution to explain how the appropriate functional structures among the inappropriate ones "get weeded out for survival" (52); the more complex those functional structures are, "the more difficult it will be to discover that they are at all appropriate" (81). Dennett thus claimed that "[l]earning can be viewed as *self-design*" (Dennett 1981b: 84).

[19] According to Dennett:

> We should reserve the term 'intelligent storage' for storage of information that is *for* the system itself, and not merely *for* the system's users or creators. For information to be *for* a system, the system must have some *use* for the information, and hence the system must have needs. The criterion for intelligent storage is then the appropriateness of the resultant behaviour to the system's needs given the stimulus conditions of the initial input and the environment in which the behaviour occurs. (46–47)

Furthermore, intra-cerebral evolution has specific constraints[20]:

> There must be conflict and something must give. Clearly what must stand firm are the inherited connections. No other conflict, and no other outcome of the conflict, would resolve itself along appropriate lines. The inherited wiring or programming must be granted hegemony in all conflicts if the plasticity of the brain is not to undo the work of species evolution and leave the animal with no appropriate responses at all […] Many animals are born with mature capacities for locomotion and discrimination of objects in their environment, but the greater the initial ability, the more rigid the brain, and hence the less adaptable the animal. More intelligent animals require longer periods of infancy, but gain in ability to cope with novel stimuli because of the higher proportion of 'soft' programming – programming not initially wired in and hence more easily overruled by novel stimuli. (57, 59)[21]

Thus we may ask: "how does the pre-established 'significance' of *some* afferent impulses allow the brain of a learning organism to discriminate appropriately the *other* impulses, which are not genetically endowed with any 'significance'?" (51).

On the basis of Cartesian metaphysical predilections, an answer to this could appeal to intra-cerebral little-men who manage learning processes. However, the less favored the belief in the existence of those little-men, the more detailed the alternative (scientifically-inspired) answer. The explanatory goal of the hypothesis of learning as brain evolution was explaining "how the brain *uses* information intelligently" (82) without postulating intra-cerebral homuncular agents – "the little man in the brain" (51) or a committee of intra-cerebral "correspondents" (87).[22]

Explaining discriminations of inputs by their environmental significance requires to account for pre-wired functional structures and the needs of the species to which the relevant system belongs, as well as to account for the brain's capacity to form new functional structures. In short, understanding the selection of species as well as the selection of functional structures is required for understanding why and how systems make discriminations by significance.

Broadly speaking, a fully-fledged account of the correctness conditions of content ascriptions requires understanding learning and intelligence on the basis of scientifically respectable hypotheses.[23] The intentional stance (strictly speaking, its

---

[20] It is worth mentioning that in *C&C* the embodied brain is characterized as endowed with (genetically transmitted) overruling pre-wired functional structures (62–63) giving rise to tropisms (like food-seeking) and action reflexes (71), as well as with the capacity to produce compound afferent-efferent functional structures which "could be 'rebuilt' piecemeal under certain conditions" (56).

[21] In the Chap. 8 of this volume Fridland argues that *C&C* advances the articulation of a framework involving a strong conceptual link between learning and intelligent storage and use of information.

[22] Dennett later claimed: "[in *C&C*] I scorned theories that replaced the little man in the brain with a committee. This was a big mistake, for this is just how one gets to 'pay back' the 'intelligence loans' of intentionalist theories" (Dennett 1981b: 81).

[23] The implicit link between each bit of Intentional interpretation and its extensional foundation is a hypothesis or series of hypotheses *describing the evolutionary source of the fortuitously propitious arrangement in virtue of which the system's operation in this instance makes sense*. These hypotheses are required in principle to account for the appropriateness which is presupposed by the Intentional interpretation, but which requires a genealogy from the standpoint of the extensional, physical theory. (80. Italics mine)

forerunner in *C&C*), evolutionary naturalism and the hypothesis of learning as intra-cerebral evolution were underpinned by a theory of content according to which content ascriptions are accurate insofar as the relevant system makes certain discriminations by significance. Thence, the afferent part of functional structures can be legitimately characterized as indicating a message or that something signifies something *for* the relevant system. In the next section I sketch the main tenets of this early version of a teleofunctional theory of content.

1.1.4 The First Teleofunctional Theory of Content

This seed was also firmly planted by Dennett on the basis of the forerunner of the intentional stance and the evolutionary considerations presented above. *C&C*'s theory of content was intended to answer the following questions: what are the vehicles of mental content? What are the empirical grounds of correctness for content ascriptions?

According to what has been said, "sense has been made of the […] claim that certain types of physical entities are systems such that their operations are *naturally* to be described in the Intentional mode – and this, only in virtue ultimately of their physical organization" (89).

Dennett was strongly influenced by Putnam's Turing-machine functionalism (Putnam 1967):

> the Intentional characterization of an event or state – identifying it, that is, as the event or state having a certain content – fixes its identity in almost the same way as a machine-table description fixes the identity of a logical state. The difference is that an Intentional characterization *only alludes to or suggests* what a machine-table characterization determines completely: the further succession of states. (112. Italics mine)

Dennett favored the view that considerations about multiple realizability of functional states should be framed by considerations about physical structures. As Ross claims: "[w]e cannot seriously take the intentional stance toward a rock or an electron because the facts of the matter in these cases will not support our doing so" (Ross 2000: 19). In Dennett's words:

> Content is a function of function […] but not every structure can realize every function, can reliably guarantee the normal relationship required. So function is a function of structure. There are, then, strong indirect structural constraints on things that can be endowed with content. If our brains were as homogeneous as jelly we could not think. (Dennett 1981h: 106)

Furthermore, the relevant phenomena are legitimately characterized as intentional "in virtue of being phenomena of goal-directed information processing systems" (ibíd.). In short, accounting for content ascriptions "runs" hand-in-hand with a revised notion of intentionality as well as with physical and functional characterizations of information-processing systems.

As it has been said, ascribing content consists in relating intentional sentences to certain phenomena – not because these sentences are ultimate vehicles of meaning, but rather because "they have meaning only in so far as they are the ploys of ultimately non-linguistic systems" (88). So the information conveyed by intentional sentences "is not preserved like a fossil in a rock; [rather,] a sentence is a vehicle for information only in that it is part of a system that necessarily includes sub-systems that process, store and transmit information non-linguistically" (ibíd.).

A theory of mental content provides a full-fledged account of the relations between intentional descriptions and extensionally described phenomena (83). Crucially, the explanatory role of intentional characterizations indeed differs from that of extensional descriptions:

> The content one ascribes to an event, state or structure is not, then, an extra feature that one *discovers* in it, a feature which, along with its other, extensionally characterized features, allows one to make predictions. Rather, the relation between Intentional descriptions of events, states or structures (as signals that carry certain messages or memory traces with certain contents) and extensional descriptions of them is one of *further interpretation*. (78)

Moreover,

> If one does ascribe content to events, the system of ascription in no way interferes with whatever physical theory of function one has at the extensional level, and in this respect endowing events with content is like giving an interpretation to a formal mathematical calculus or axiom system, a move which does not affect its functions or implications but may improve intuitive understanding of the system. (79)[24]

Dennett claimed that we cannot know whether a system, *s*, discriminates, e.g., a square *as* a square (or, in other words, whether the stimulus condition *C* signifies 'square' for *s*) at *t* merely by knowing the afferent phenomena taking place in *s* at *t* when presented with – what, *for us* (observers or users of *s*) is – a square (74, ff.). In spite of knowing the afferent patterns triggered by *C* in *s* at *t*, the question whether *C* signifies 'square' (or, e.g., 'warning') for *s* at *t* remains open, because one does not know the efferent phenomena also indirectly propitiated by *C* in *s* at *t* (79, ff.). So Dennett claimed:

> There should be possible some scientific story about synapses, electrical potentials and so forth that would explain, describe and predict all that goes on in the nervous system. If we had such a story we would have in one sense an extensional theory of behaviour, for all the *motions* (extensionally characterized) of the animal caused by the activity of the nervous system would be explicable and predictable in these extensional terms, but one thing such a story would say nothing about was *what the animal was doing*. This latter story can only be told in Intentional terms, but it is not a story about features of the world *in addition to* the features of the extensional story; it just describes what happens in a different way. (78)

[24] In this way, Dennett claims: "I certainly am *not* aware of […] neural activities, while I *am* aware of my thoughts […] in any event the *content* of the [neural] activities is not at all a discriminable characteristic of them […] but merely an artificial determination made by some observing neurologist" (107).

Dennett thus claimed that specifying the content of an afferent phenomenon requires knowing certain behavioral responses of the relevant system. The teleofunctional[25] theory of content was sketched in *C&C*[26] along the following lines:

> No afferent can be said to have the significance 'A' until it is 'taken' to have the significance 'A' by the efferent side of the brain, which means, unmetaphorically, until the efferent side of the brain has produced a response (or laid down response controls) the unimpeded function of which would be appropriate to having been stimulated by an A […] what an event or state 'means to' an organism also depends on what it *does* with the event or state. (74, 76)

The core claim of this theory is that ascribing content requires knowledge of information processing afferent-efferent phenomena:

> [n]o physical motions or events have intrinsic significance. [For instance, the] electrical characteristics of an impulse sequence, or the molecular characteristics of a nerve fibre could not independently determine what the impulses *mean*, or what *message* the nerve fibre carries, and therefore what a stimulus – however complex – heralds cannot be a function of its internal characteristics alone. (47)

In this way, "the [explanatory] shift from [reference to the environmental significance of afferent phenomena] to an object reference must depend on what effect [the inputs have] on behaviour" (83). For instance, "no structure or state could be endowed with the storage content 'thin ice is dangerous', no matter how it had been produced, if the input 'this is thin ice' did not cause it to produce an appropriate continuation, such as 'do not walk on the ice'" (84).

In order to specify the "vehicles" of content, intentional characterizations should be supported by knowledge of functional and physical states of the relevant system. In this way, this theory of content requires an account of the origins of the functional structures that enable appropriate afferent-efferent sequences (73). This can be brought out by example:

> Suppose that in an organism O there is a particular highly interpreted afferent output A (summing, we can suppose, signals from visual, tactile and olfactory sources) that fired normally if and only if food was present in O's perceptual field. The firing of A might have any of a vast number of effects on O's behaviour. If it happened for example to have the effect of terminating a series of 'seeking' sub-routines and initiating a series of other, 'eating' sub-routines, we would have evidence for saying that O has achieved its goal of finding food, has recognized that the goal was achieved, had discriminated the presence of food *as* the presence of food. If, on the other hand, A did not have this effect, if O did not commence eating or in other ways behave appropriately to the presence of food under the circumstances, then regardless of any evidence we might have about the specificity of the stimulus conditions determining the firing of A, there would be no reason to say that the animal had discriminated the presence of food *as* the presence of food. (73)

---

[25] In *C&C* 'centralism' is the closest label indicating the systematic set of ideas that later became the pillar of Dennett's teleofunctionalism (see: 83–86).

[26] Three decades after *C&C*, Dennett proclaimed himself as "the original teleofunctionalist (in *Content and Consciousness*)" holding that he didn't make "the mistake of trying to define all salient mental differences in terms of biological functions. That would be to misread Darwin badly" (1991: 460). Here Dennett seems to be making reference to Millikan's teleosemantics (Millikan 1984). For some remarks about the difference between Dennett's and Millikan's works, see: Ross (2000: 11–12).

In short, "[t]he content, if any, of a neural state, event or structure depends on two factors: its normal source in stimulation, and whatever *appropriate* further efferent effects it has; and to determine these factors one must make an assessment that goes beyond an extensional description of stimulation and response locomotion" (76). Such "assessment" depends on adopting the intentional stance.

The significance of an environmental entity must be specified by adopting the intentional stance on the basis of knowledge of the functional and physical structures of the relevant system, its environmental situation and certain conditions by which its species was favored by natural selection.

As it has been suggested, understanding systems' activities by adopting the intentional stance don't overextend the catalog of entities described in purely extensional terms, because only the extensional vocabulary is referential. Further, understanding the relations between the extensional and the intentional vocabulary doesn't require a reductive analysis of the latter. Thus, it seems that the intentional vocabulary has a sort of autonomy with respect to the extensional vocabulary. In *C&C*, Dennett introduced the personal/sub-personal distinction in order to clarify the scope and function of the intentional vocabulary, as well as its relations to the extensional explanatory domain.

1.1.5 The Personal/Sub-personal Distinction

The personal/sub-personal distinction is fundamental in order to understand and support the idea that the intentional vocabulary is not referential, despite the fact that it is required for explaining mind and behavior in intelligent systems.

The personal level and the sub-personal level differ deeply regarding the subject matter (149). "People can reason, but brains cannot, any more than feet (or whole bodies) can flee or a hand can sign a contract. People can *use* their feet in fleeing or their hands in signing a contract, but it would not be correct to say in the same sense that people use their brains in thinking and reasoning" (149).[27]

Dennett thus claimed:

> The recognition that there are two levels of explanation gives birth to the burden of relating them, and this is a task that is not outside the philosopher's province. It cannot be the case that there is *no* relation between pains and neural impulses or between beliefs and neural states, so setting the mechanical or physical questions off-limits to the philosopher will not keep the question of what these relations are from arising. The position that pains and beliefs are in one category or domain of inquiry while neural events and states are in another cannot be used to isolate the philosophical from the mechanical questions, for […] different

[27] The personal/sub-personal distinction has been widely discussed (see: Elton 2000; Hornsby 2000; Bermúdez 2000; Davies 2000). There's agreement with respect to the seminal role that the distinction (and its reformulations) has(ve) played in philosophy of mind, cognitive science, cognitive psychology, and related fields. For critical reviews, see: Skidelsky (2006) and Drayson (2014). In this volume Frankish, Wilkinson, and Roth (Chaps. 4, 6, and 7 respectively) develop detailed accounts about the distinction. Hornsby (2000) claims that during the 1970s Dennett re-formulated the distinction introduced in *C&C*. Roth grants a similar view. See*:* Dennett (1987).

> categories are no better than different Cartesian substances unless they are construed as different ontological categories, which is to say: the *terms* are construed to be in different categories and only one category of terms is referential. (95)

Entering the sub-personal level requires to abandon the level in which the notion of 'person' is operative, because at the sub-personal level persons – as whole system with desires, needs, beliefs, and so on – do not play any causal explanatory role. Sub-personal explanations thus concern the extensional grounds of intentionally described personal activities. However, this doesn't require abandoning personal explanations because that would also lead to abandon the explanatory level required to explain mental phenomena. For instance, in the case of pain,

> [W]hen we abandon mental process talk for physical process talk we cannot say that the mental process analysis of *pain* is wrong, for our alternative analysis cannot be an analysis of pain at all, but rather of something else – the motions of human bodies or the organization of the nervous system. Indeed, the mental process analysis is correct. Pains are feelings, felt by people. (94)

Dennett argued that the "only way to foster the proper separation between the two levels of explanation, to prevent the contamination of the physical story with unanalysable qualities or 'emergent phenomena', is to put the fusion barrier between them" (96). That "fusion barrier" would avoid identifications, for instance, of thoughts, desires, sensations and beliefs "with anything in the sub-personal story" (113).

The intentional vocabulary is not restricted to the personal level, since (depending on certain explanatory needs) intentional notions are required for making sense of some sub-personal information-processing events.[28] The "fusion barrier" must be preserved in order to avoid the categorical error of granting that intentional characterizations of sub-personal systems are on a par – at the same explanatory level – with their extensional descriptions.[29]

Dennett applied the views, hypotheses, concepts and distinctions that have been summarized above in order to account for particular phenomena, like introspective certainty and mental imagery. In the next section I'll sketch his view about introspective certainty and in Sect. 1.1.7. about mental imagery.

1.1.6 Explaining Introspective Certainty in Terms of Logical States of the Cerebral Computer

According to the Cartesian view, introspection is the mental act by which one consciously accesses one's mind. According to this classical view, introspective reports (like "I think that I forgot her mobile number") convey information about private

[28] For instance, in the case of the AI researcher: he "*starts* with an intentionally characterized problem (e.g., how do I get the computer to *recognize* questions, *distinguish* subjects from predicates, *ignore* irrelevant parsings?) and then breaks these problems down still further until finally he reaches problem or task descriptions that are obviously mechanistic" (Dennett 1981b: 80).

[29] As Fodor's intentional realism suggests. See*:* Fodor (1975). For a reply, see: Dennett (1981h).

mental contents. This view includes the claim that introspection is a mode by which persons access their minds and the claim that introspective reports are somehow infallible and incorrigible.

In *C&C* Dennett developed a (Putnamian) functionalist explanation of accessibility and incorrigibility. This explanation was further developed and corrected by him some decades after *C&C* (see, e.g., Dennett 2000, 2002, 2007). Nonetheless, in the "Second Preface" of *C&C* (xi), Dennett claimed that there are some salvageable points, in particular, about the identity conditions of mental states and their relations to reports about them.

Dennett remarked that explaining introspective access and introspective incorrigibility requires a sub-personal account, because

> On the personal, mental language level we still have a variety of dead-end truths, such as the truth that people just *can* tell what they are thinking, and the truth that what they report are their thoughts. These are truths that deserve to be fused, and then the fact that there should be such truths can be explained at another level, where people, thoughts, experiences and introspective reports are simply not part of the subject matter. (113)

C&C provided a sub-personal description of the linguistic behavior controls enabling the production of introspective reports (107). Accordingly, "the immunity to error [of introspective reports] has nothing to do with the execution of any *personal action* […] an account of a man's intention […] plays no role in explaining introspective certainty" (111). According to his view, for instance, introspective reports about perceptual experiences are described like outputs of the speech center, whose inputs are outputs of sensory analysers, such that there might be two sources of error: one in sensory analysers and other in the speech center. However, if they are working properly, the speech center "cannot misidentify the output which comes from the [sensory] analyser, which is the same logical state as the speech centre input" (110).

According to this account, introspective reports do not *refer* to informational processes in the cerebral computer; they rather *assign* content to sensory analysers' outputs. "There is no entity […] in the human brain […] that would be well *referred* to by the expression 'that which is infallibly reported by the final output expression,' and this is the very best of reasons for viewing this expression and its mate, 'thought', as non-referential" (113). In short, the information conveyed by introspective reports is not about a part of the system.

The content assigned by an introspective report is *what-is-reported* (e.g., beliefs, desires and pains). Further, a cerebral computer's state (112) is *what-is-expressed*. Consequently, "[s]tarting from the position that thoughts, being what-is-reported, cannot be identified with anything in the sub-personal story, it would be poor philosophy to argue further that there must really *be* something, the thought, that is reported when it is true that I am reporting my thoughts" (113).

"A Turing machine designed so that its output could be interpreted as reports of its logical states[30] would be, like human introspectors, invulnerable to all but 'ver-

[30] "A particular machine T is in logical state A if, and only if, it performs what the machine table specifies for logical state A, regardless of the physical state it is in" (102). Dennett's neurocomputational account was clearly influenced by Putnam's Turing-machine functionalism (Putnam 1967).

bal' errors. It could not misidentify its logical states in its reports just because it does not have to identify its states at all" (103–104). If introspective reports neither refer to nor identify any logical states of the cerebral computer, they strictly speaking don't misidentify anything in the system – they cannot be fallible with respect to the states of the relevant system (110). This description amounts to the view that introspective reports only have expressive value with respect to the states of the cerebral computer.

After *C&C* Dennett wrote:

> if we say what we mean to say, if we have committed no errors or infelicities of expression, then our actual utterances cannot fail to be expressions of the content of our semantic intentions, cannot fail to do justice to the access we have to our own inner lives […] I claimed in *Content and Consciousness* that this fact explained how we were, in a very limited and strained sense, incorrigible with regard to the contents of our awareness or consciousness. Now, thanks to the relentless persuasions of John Bender, William Talbott, Thomas Blackburn, Annette Baier and others, I wish to claim that this fact explains not how we are in fact incorrigible, but rather why people – especially philosophers – so often think we are. (Dennett 1981g: 171)

1.1.7 A Pioneering Critique of Pictorial Models of Imagination

C&C provided the first critique against the pictorial doctrine of mental imagery, which was followed by a cascade of further attacks (see: Pylyshyn 1973). Later on, a heated debate propitiated the development of sophisticated versions of that doctrine (see: Kosslyn 1975, 1976).

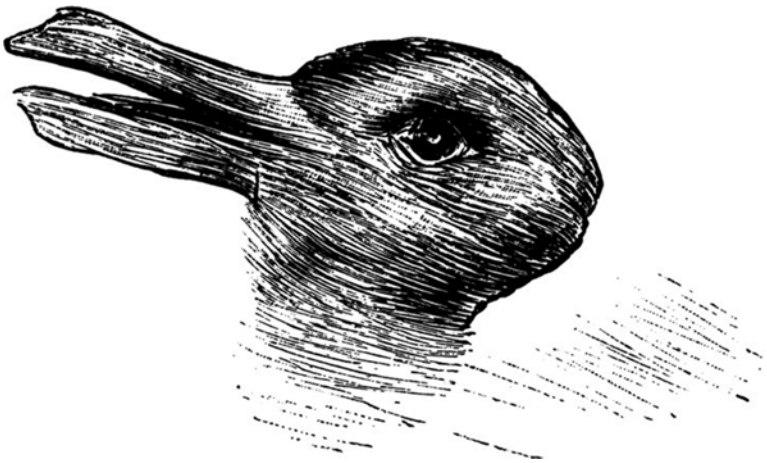
According to the pictorial doctrine, perceiving, remembering a visual episode and imagining a visual scenario are ways of being conscious of mental-items that resemble other items by their shape, form and color.[31] Dennett focused his analysis on visual perception and argued that in it there are not elements which represent in virtue of resembling what they represent (135ff.). In short, Dennett claimed that visual consciousness is not filled with mental pictures.

From a sub-personal account of mental imagery there is no place for images among the functional structures realizing perceptual processes. The main reason for this is that, at the sub-personal level, images can work as images only if there is a person "(or an analogue of a person) to see or observe it, to recognize or ascertain the qualities in virtue of which it is an image of something" (134). In other words, at the sub-personal level persons are not causally operative, so there cannot be anything working as an image at that level.

Furthermore, from a personal account, mental images appear to be descriptions rather than pictorial representations of what they represent: when we see something

[31] The pictorial doctrine was widely endorsed by Modern philosophers, like Descartes, Locke, Hume, Berkeley, Reid, and Kant, and also by contemporary philosophers, like Russell, Meinong, and C. Lewis. During the 1960s, in experimental psychology, the doctrine was defended, e.g. see: by Shepard (1966), Bahrick and Boucher (1968), and Bugelski (1968).

we are aware of "an edited commentary on the things of interest", not of a fine-grained picture of those things (136). Look at the following famous picture[32]:



Dennett explained this case as follows:

> The image (on the paper or the retina) does not change, but there can be more than one description of that image […] One says at the personal level 'First I was aware of it *as* a rabbit, and then *as* a duck', but if the question is asked 'What is the difference between the two experiences?', one can only answer at this level by repeating one's original remark. To get to other more enlightening answers to the question one must resort to the sub-personal level, and here the answer will invoke no images beyond the unchanging image on the retina. (137)

Like in this cases, Dennett claimed that the philosophical problem about visual hallucinations also vanishes insofar as we stop thinking of visual perception as inner scanning of mental pictures.

Dennett claimed that his account in *C&C* should be viewed as a contribution "(good or bad) to psychology, not to philosophy" (1981i: 189). Later on, he explicitly discredited "spurious […] debates about entirely mythical species of mental images: the various non-physical, phenomenal or epiphenomenal, self-intimating, transparent to cognition, unmisapprehensible, pseudo-extended, quasi-imagistic phantasms that have often been presented as mental images in the past" (ibíd., 188. See also Dennett 1991).

Every seed required additional conceptual "engineering" for growing up after the 1970s. All of them outgrew the boundaries of philosophical gardens.

[32] "Kaninchen und Ente" ("Rabbit and Duck"). In *Fliegende Blätter*, (Oct. 23, 1892, 147). See: http://diglit.ub.uni-heidelberg.de/diglit/fb97/0147&ui_lang=eng

1.2 A Sea of Prolific Works

C&C developed a meta-theoretical plan for understanding the multilayered intersections between neurophysiology, cognitive psychology, evolutionary theory, computer science and analytic philosophy. In short, it showed that developing a science of consciousness is part of the philosophical agenda. Indeed, *C&C* was one of the first accounts of consciousness and mental content from an interdisciplinary approach offered within the analytic philosophy tradition. Some decades after *C&C*, inter-disciplinarity has been taken as a requirement for theorizing on mental content and consciousness.[33]

It's worth mentioning that during the second half of the nineteenth century and the first decade of the twentieth century philosophers, physiologists and psychologists also purported to consolidate empirically-supported frameworks for explaining consciousness, perception, and voluntary actions. For instance: psychophysics,[34] American functionalist psychology[35] and comparative evolutionary psychology.[36] These naturalist frameworks rejected the development of a purely intentional science ("psychognosy"[37]); like *C&C* 60 years later purported to do (see Sect. 1.1.1).

Several scientific accounts of mind and behavior at the end of the nineteenth century and in the first half of the twentieth century still preserved Cartesian assumptions; although, e.g., some anti-Cartesian neo-Kantian efforts (see: Vogt 1847; Büchner 1855) or the naturalist core of the so called psychophysical parallelism endorsed by some materialists and neurobiologists (see: Lange 1873; Sherrington 1947).

---

[33] See, e.g.: Gazzaniga and LeDoux (1978), Dretske (1981), Millikan (1984), Minsky (1986), Lycan (1987), Jackendoff (1987), Baars (1988), Penrose (1989), Edelman (1989), Dennett (1991), McGinn (1991), Humphrey (1992), Flanagan (1992), Churchland and Sejnowski (1992), Crick (1994), Pinker (1994, 1997), Clark (1997), Ramachandran and Blakeslee (1998), Block (2001), Llinas (2001), Prinz (2005), Gallagher (2005), Carruthers (2006), Tye (2009), Burge (2010), Damasio (2010), Tononi (2012). For a general view on the interdisciplinary debate on consciousness, See: Freeman (2003).

[34] See, e.g., Weber (1851), Fechner (1860), von Helmholtz (1863, 1867), Mach (1866), Wundt (1871).

[35] See, e.g., James (1890), Thorndike (1905, 1911), Yerkes (1907, 1911).

[36] See, e.g., Spencer (1855), Darwin (1872), Huxley (1873, 1898), Morgan (1894), Romanes (1882), Hobhouse (1901).

[37] For instance, as presented by Brentano:

> Psychognosy is different. It teaches nothing about the causes that give rise to human consciousness and which are responsible for the fact that a specific phenomenon does occur now, or does not occur now or disappears. Its aim is nothing other than to provide us with a general conception of the entire realm of human consciousness. It does this by listing fully the basic components out of which everything internally perceived by humans is composed, and by enumerating the ways in which these components can be connected. Psychognosy will therefore, even in its highest state of perfection, never mention a physico-chemical process in any of its doctrines [*Lehrsatz*]. (Brentano 2002: 3–4)

Right after its publication, C&C was tackled from several flanks. Smart (1970) discussed C&C's plan by asking to what extent intentional statements can be true or false if they are taken to be non-referential. Gundersen (1972) attacked the significance of the personal/sub-personal distinction by arguing that it leads to a recycled version of the mind-body problem. Nagel (1972) claimed that *C&C* leaves the mind-body problem "undisturbed" because of failing to explain consciousness and, in particular, because it confuses a genuine explanation of consciousness with descriptions of things that are compatible with the absence of consciousness (e.g., behavioral patterns).[38] *C&C* was born in a fruitful research environment in which several new paradigms in the study of mind and behavior emerged.

C&C took part of in an anti-behaviorist wave[39] that appeared in the vortex of seminal advances in the scientific study of mind and behavior. Between the 1940s and the 1970s not only new theories[40] and approaches emerged, but also new fields of study, like cognitive psychology.[41] The late 1960s and the early 1970s witnessed three major contributions to the interdisciplinary plan: general systems theory (von Bertalanffy 1968), the system dynamics theory (Forrester 1971), and second-order cybernetics (von Foerster 1974).

Those fields derived from the joint work of engineers, psychologists, neurophysiologists, mathematicians, and philosophers who believed that behavior and intelligence could be explained by adopting a computational and mathematical approach. Several problems of deep philosophical significance, such as the "frame problem"[42] (McCarthy and Hayes 1969) were discovered in this way. Formal models of neural functioning (see, e.g., Hebb 1949; Arbib 1964) and findings in neural circuitry were simultaneously taking place.[43] The philosophical labor required "distilling" a scientifically respectable conception of mind and behavior.[44] *C&C* is one of the seminal works in analytic philosophy that carried on that labor.

[38] For more reactions, see: Blake (1969), Dent (1970), Franklin (1970), Kane (1970), McKim (1970), Rice (1971), Arbib (1972), Audi (1972).

[39] See: Miller (1956), Bruner (1966), Newell et al. (1958), Chomsky (1959), Neisser (1963), Putnam (1964, 1967a, b), Taylor (1964), Fodor (1968). See, e.g., Efron (1967).

[40] E.g., information theory (see, e.g., Shannon 1948; von Neumann 1955), cybernetic theories (see, e.g., Rosenblueth et al. 1943; Wiener 1948; Ashby 1952, 1956) and artificial intelligence theories (see, e.g., McCulloch and Pitts 1943; von Neumann 1945, 1951, 1958; Shannon 1948, 1950a, b; Turing 1948, 1950; McCarthy et al. 1955; Newell and Simon 1963; Minsky 1967).

[41] See, e.g., Miller (1956), Chomsky (1959), Bruner (1966), Neisser (1967, 1976).

[42] See: Dennett (1981f, 1984).

[43] See, e.g., Walter (1950a, b, 1951, 1953), Young (1964, 1965), Penfield and Rasmussen (1950), Lettvin et al. (1959), Ratliff and Hartline (1959), Hubel and Wiesel (1962).

[44] See, e.g., Place (1956), Feigl (1958), Smart (1959), Putnam (1960, 1964, 1967).

1.3 On This Book

This book provides fresh views about the foundations of the theoretical system sketched in *C&C*. The chapters cover the fundamental concepts, hypotheses and approaches introduced in *C&C*, taking into account the findings and progress that have taken place during more than four decades. This volume is a multi-authored revisited version of *C&C*.

Don Ross (Chap. 2) argues that *C&C* contains a scientific discovery indicating an (incomplete) unification of empirical findings. Ross explains why this discovery can be seen as a "rare achievement" by a philosopher and reflects in detail about the contributions that philosophical theorizing can make to our scientific understanding of the mind. As a result, Ross articulates a view on the relations between science and philosophy and offers an interpretation of *C&C*.

Felipe De Brigard (Chap. 3) argues against both eliminative materialism and Fodorian intentional realism about propositional attitudes by developing an anti-realist interpretation of propositional attitude ascriptions. Even if a propositional attitude ascription is true, De Brigard argues, it doesn't follow that there is a sentence in the language of thought corresponding to the "that" clause in the ascription. After he developed this view, De Brigard was surprised to discover that *C&C* already defended the same view, albeit via a different route.

Keith Frankish (Chap. 4) introduces and defends a reinterpretation of dual-process theories of reasoning on the basis of a systematic review of *C&C*'s remarks about thinking, awareness and the personal/sub-personal distinction. Frankish associates this reinterpretation with the "dual-attitude theory of belief" and concludes that psychological theories of reasoning have neglected the personal/sub-personal distinction.

Richard Dub (Chap. 5) inquires about Dennett's motivations and the explanatory role played by the claim that attributing cognitive states entails that the relevant system is rational. This became a core claim in Dennett's theory of intentional systems.

Sam Wilkinson (Chap. 6) clarifies two versions of Dennett's personal/sub-personal distinction by introducing considerations about their explanatory role in cognitive neuropsychiatry. Wilkinson argues that the distinction introduced in *C&C* can be used to describe and predict the behavior of subjects with mental disorders within the perspective of cognitive neuropsychiatry.

Martin Roth (Chap. 7) interprets the "extended mind hypothesis" on the basis of considerations about the personal/sub-personal distinction introduced by Dennett's and Fodor's early works. Roth claims that endorsing the "extended mind hypothesis" requires specifying the explanatory role of sub-personal intentional explanations and elucidates the usages of the distinction that would derive either from endorsing or from rejecting the "extended mind hypothesis."

Taking *C&C* as a point of departure, Ellen Fridland (Chap. 8) focuses on the claim that learning is intimately related to intelligent information processing. She concludes that the presence of past or future learning is necessary to qualify a

behavior or mental process as intelligent. She aims at showing that a detailed inquiry about the features of intelligence (e.g., flexibility, transferability, manipulability, and appropriateness) should be framed in terms of learning.

John Michael (Chap. 9) doesn't develop a direct interpretation of some hypothesis introduced in *C&C*. Michael's chapter rather purports to show that the intentional stance can be articulated as a platform for introducing a new theoretical framework accounting for the psychological development of cultural learning. Michael develops the view that cultural learning consists in a "feedback loop" involving the multi-personal adoption of the intentional stance and concludes that this view helps to explain the reliability of the intentional stance.

Pete Mandik (Chap. 10) focuses on a thesis suggested in *C&C* but later detailed in Dennett's *Consciousness Explained* (i.e. Dennett's "first-person operationalism"). His arguments press on the higher-order thought theorists of consciousness by introducing a dilemma: either they accept a relational interpretation of the higher-order thought theory and develop a satisfactory reply to Mandik's "Unicorn Argument", or they recognize that the higher-order thought theory collapses into first-person operationalism (Denettian anti-realism).

Finally, with his distinctive constructive, critical, and creative style, Dennett (Chap. 11) brings out the pros and cons of the other chapters as well as their innovations.

The philosophical-scientific vegetation will spread more seeds. Some will be firmly planted; others won't germinate; yet others will require further engineering to germinate. At the end of the day, the lay person and the expert reader will stare at the landscape and decide either to chill out in the gardens they already know, or to pass through the jungle in search of unexplored fields.[45]

References

Arbib, M. (1964). *Brains. Machines and the mathematics*. New York: McGraw Hill.

Arbib, M. (1972). Consciousness: The secondary language. *The Journal of Philosophy, 69*(18), 579–591.

Ashby, W. (1952). *Design for a brain: The origin of adaptive behavior*. New York: Chapman and Hall.

Ashby, W. (1956). *An introduction to cybernetics*. New York: Wiley.

Audi, R. (1972). Review to *content and consciousness. Philosophy Forum, 12*, 206–208.

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.

Bahrick, H., & Boucher, B. (1968). Retention of visual and verbal codes of the same stimuli. *Journal of Experimental Psychology, 78*, 417–422.

Bermúdez, J. (2000). Personal and sub-personal. A difference without a distinction. *Philosophical Explorations, 31*, 63–82.

Blake, A. (1969). Review to content and consciousness. *Systematics, 7*, 261–263.

[45] I am grateful to Felipe De Brigard, Gualtiero Piccinini and John Horden for their comments and suggestions.

Block, N. (2001). How not to find the neural correlate of consciousness. In J. Branquinho (Ed.), *The foundations of cognitive science* (pp. 1–10). Oxford: Clarendon.

Brentano, F. (2002). *Descriptive psychology*. Londres/Nueva York: Routledge. Originally: Chisholm, R., & Baumgartner, W. (Eds.) (1982). *Deskriptive psychologie*. Hamburg: Meiner.

Brentano, F. (2009). *Psychology from an empirical standpoint*. London/New York: Routledge. Originally published: 1874. *Psychologie vom empirischen Standpunkt*. Leipzig: Duncker and Humblot.

Bruner, J. (1966). *Studies in the cognitive growth*. New York: Wiley.

Büchner, L. (1855). *Kraft und stoff* (5th ed.). Frankfurt am Main: Meidinger.

Bugelski, B. (1968). Images as mediators in one-trial paired-associate learning. II: Self-timing in successive lists. *Journal of Experimental Psychology, 77*, 328–334.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought*. New York: Oxford University Press.

Changeux, J.-P., & Danchin, A. (1976). Selective stabilization of developing synapses as a mechanism for the specification of neuronal networks. *Nature, 264*, 705–712.

Changeux, J.-P., & Dehaene, S. (1989). Neural models of cognitive functions. *Cognition, 33*, 63–109.

Changeux, J.-P., Courrège, P., & Danchin, A. (1973). A theory of the epigenesis of neural networks by selective stabilization of synapses. *Proceedings of the National Academy of Sciences of the United States of America, 70*, 2974–2978.

Chisholm, R. (1957). *Perceiving: A philosophical study*. Ithaca: Cornell University Press.

Chomsky, N. (1959). A review of B. F. Skinner verbal behavior. *Language, 35*(1), 26–58.

Churchland, P., & Sejnowski, T. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there. Putting brain. Body and world together again*. Cambridge, MA: MIT Press.

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. New York: Scribners.

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. New York: Pantheon Books.

Darwin, C. (1872). *The expression of emotion in man and animals*. London: John Murray.

Davies, M. (2000). Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind and Society, 2*(1), 87–105.

Dennett, D. (1981a). *Brainstorms. Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Dennett, D. (1981b). Why the law of effect will not go away. In Denett 1981: 70–89. Originally published: 1975. *Journal of the Theory of Social Behaviour*, *2*, 169–187.

Dennett, D. (1981c). Mechanism and responsability. In Denett 1981a: 233–255. Originally published: Honderich, T. (Ed.) (1973). *Essays on freedom of action*. London/Boston: Routledge and Kegan Paul.

Dennett, D. (1981d). Intentional systems. In Denett 1981a: 3–22. Originally published: 1971. *Journal of Philosophy LXVIII*(4), 87–106.

Dennett, D. (1981e). Reply to Arbib and Gunderson. In Dennett 1981a: 23–38. Originally published: 1972. *Journal of Philosophy LXIX*(18): 604.

Dennett, D. (1981f). Artificial intelligence as philosophy and as psychology. In Dennett 1981a: 108–126. Originally published: Martin Ringle (Ed.) (1978). *Philosophical perspectives on artificial intelligence*. New York: Humanities Press and Harvester press.

Dennett, D. (1981g). Toward a cognitive theory of consciousness. In Denett 1981a: 149–173. Originally published: Martin Ringle (Ed.) (1978). *Philosophical perspectives on artificial intelligence*. New York: Humanities Press and Harvester press.

Dennett, D. (1981h). A cure for the common code? In Denett 1981a: 90–108. Originally published: Fodor, J. (1977). Critical notice: The language of thought. *Mind*, *86*: 265–280

Dennett, D. (1981i). Two approaches to mental imagery. In Denett 1981a: 174–189.

Dennett, D. (1984). Cognitive wheels: The frame problem in artificial intelligence. In C. Hookway (Ed.), *Minds, machines, and evolution: Philosophical studies* (pp. 129–151). Cambridge, UK: Cambridge University Press.
Dennett, D. (1986). *Content and consciousness* (2nd ed.). London/Boston/Henley: Routledge and Kegan Paul. Originally published: 1969. London: Routledge and Kegan Paul.
Dennett, D. (1987). *The intentional stance*. Cambridge, MA: Bradford Books/MIT Press.
Dennett, D. (1991). *Consciousness explained*. Boston: Little Brown.
Dennett, D. (1995). *Darwin's dangerous idea: Evolution and the meaning of life*. New York: Simon and Schuster.
Dennett, D. (2000). The case for rorts. In R. Brandom (Ed.), *Rorty and his critics* (pp. 91–101). Malden: Blackwell.
Dennett, D. (2002). How could I be wrong? How wrong could I be? *Journal of Consciousness Studies, 9*(5–6), 13–16.
Dennett, D. (2007). Heterophenomenology reconsidered. *Phenomenology and the Cognitive Sciences, 6*, 247–270.
Dennett, D. (2008). Daniel Dennett: Autobiography. Part 1. *Philosophy Now*, 68. https://philosophynow.org/issues/68/Daniel_Dennett_Autobiography_Part_1. Accessed 4 Aug 2013.
Dent, N. (1970). Review to *content and consciousness. Philosophical Quarterly, 20*, 403–404.
Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass, 9*(5), 338–346.
Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
Edelman, G. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.
Edelman, G. (1989). *The remembered present: A biological theory of consciousness*. New York: Basic Books.
Efron, R. (1967). The duration of the present. *Proceedings of the New York Academy of Science, 8*, 542–543.
Elton, M. (2000). Consciousness: Only at the personal level. *Philosophical Explorations, 31*, 25–42.
Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig: Breitkopf und Härtel.
Feigl, H. (1958). The 'mental' and the 'physical'. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories, and the mind-body problem. Minnesota studies in the philosophy of science* (pp. 370–497). Minneapolis: University of Minnesota Press.
Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press.
Fodor, J. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House.
Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
Forrester, J. (1971). *World dynamics*. Cambridge, MA: Wright-Allen Press.
Franklin, R. (1970). Review to *content and consciousness. Australasian Journal of Philosophy, 48*, 264–273.
Freeman, A. (2003). *Consciousness. A guide to the debates*. Santa Barbara: ABC-CLIO.
Gallagher, S. (2005). *How the body shapes the mind*. Oxford/New York: Oxford University Press.
Gazzaniga, M., & LeDoux, J. (1978). *The integrated mind*. New York: Plenum Press.
Gundersen, K. (1972). Content and consciousness and the mind-body problem. *Journal of Philosophy, 64*(5), 591–604.
Hebb, D. (1949). *The organization of behavior. A neuropsychological theory*. New York: Wiley.
Hobhouse, L. (1901). *Mind in evolution*. London: Macmillan.
Hornsby, J. (2000). Personal and sub-personal: A defence of Dennetts early distinction. *Philosophical Explorations, 31*, 6–24.
Hubel, D., & Wiesel, T. (1962). Receptive fields. Binocular interaction and functional architecture in the cats visual cortex. *The Journal of Physiology, 160*, 106–154.
Humphrey, N. (1992). *A history of the mind*. London: Chatto and Windus.
Huxley, T. (1873). *Evidence as to mans place in nature*. New York: D. Appleton and co.
Huxley, T. (1898). *Hume. With helps to the study of Berkeley: Essays*. New York: D. Appleton.

Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.
James, W. (1890). *Principles of psychology*. Cambridge, MA: Harvard University Press.
Kane, R. (1970). Review to *content and consciousness. Review of Metaphysics, 23*, 740.
Kosslyn, S. (1975). Information representation in visual images. *Cognitive Psychology, 7*, 341–370.
Kosslyn, S. (1976). Can imagery be distinguished from other forms of internal representation? Evidence from studies of information retrieval times. *Memory and Cognition, 4*, 291–297.
Lange, A. (1873). *Geschichte des Materialismus und Kritik seiner Bedeutung in der Gegenwart* (2 vols.). Frankfurt Main: Suhrkamp.
Lettvin, J., Maturana, H., McCulloch, W., & Pitts, H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE, 47*(11), 1940–1959.
Llinas, R. (2001). *I of the vortex: From neurons to self*. Cambridge, MA: MIT Press.
Lycan, W. (1987). *Consciousness*. Cambridge, MA: MIT Press.
Mach, E. (1866). *Beiträge zur Analise der Empfindungen*. Jena: Verlag von Gustav Fischer.
McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence, 4*, 463–502.
McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. Darmouth conferences, 31 Aug. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html. Accessed 4 Aug 2013.
McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.
McGinn, C. (1991). *The problem of consciousness*. Oxford: Blackwell.
McKim, V. (1970). Review to *content and consciousness. New Scholasticism, 44*, 472.
Miller, G. (1956). The magical number seven. Plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
Minsky, M. (1967). *Computation: Finite and infinite machines*. Englewood Cliffs: Prentice-Hall.
Minsky, M. (1986). *The society of mind*. New York: Simon and Schuster.
Morgan, C. (1894). *An introduction to comparative psychology*. London: Walter Scott. Ltd.
Nagel, T. (1972). Review to *content and consciousness. Journal of Philosophy, 20*(69), 220–224.
Neisser, U. (1963). Decision time without reaction time: Experiments in visual scanning. *American Journal of Psychology, 36*, 376–385.
Neisser, U. (1967). *Cognitive psychology*. New York: Appleton.
Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W. H. Freeman.
Newell, A., & Simon, H. (1963). GPS. A program that simulates human thought. In E. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 279–293). Cambridge, MA: MIT Press.
Newell, A., Shaw, J., & Simon, H. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*, 151–166.
Penfield, W., & Rasmussen, T. (1950). *The cerebral cortex of man: A clinical study of localization of function*. New York: Hafner Pub. Co.
Penrose, R. (1989). *The emperors new mind: Computers. Minds and the laws of physics*. Oxford: Oxford University Press.
Pinker, S. (1994). *The language instinct. How the mind creates language*. New York: W. Morrow and Co.
Pinker, S. (1997). *How the mind works*. New York: W. W. Norton.
Place, U. (1956). Is consciousness a brain process? *British Journal of Psychology, 47*(1), 44–50.
Prinz, J. (2005). A neurofunctional theory of consciousness. In A. Brook & K. Akins (Eds.), *Cognition and the brain. The philosophy and neuroscience movement* (pp. 381–396). Cambridge/New York: Cambridge University Press.
Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind* (pp. 57–80). New York: New York University Press.

Putnam, H. (1964). Robots: Machines or artificially created life? *Journal of Philosophy, 61*, 668–691.
Putnam, H. (1967a). Psychological predicates. In W. Capitan & D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: Pittsburgh University Press.
Putnam, H. (1967b). The mental life of some machines. In H.-N. Castañeda (Ed.), *Intentionality, minds and perception* (pp. 177–200). Detroit: Wayne State University Press.
Pylyshyn, Z. (1973). What the minds eye tells the minds brain: A critique of mental imagery. *Psychological Bulletin, 80*, 1–25.
Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.
Ramachandran, V., & Blakeslee, S. (1998). *Phantoms in the brain: Probing the mysteries of the human mind*. New York: William Morrow.
Ratliff, F., & Hartline, H. (1959). The response of limulus optic nerve fibers to patterns of illumination on the receptor mosaic. *Journal of General Physiology, 42*, 1241–1255.
Rice, L. (1971). Review to *content and consciousness. Modern Schoolman, 48*, 177–178.
Romanes, G. (1882). *Animal intelligence*. London: Kegan Paul. Trench and Co.
Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science, 10*, 18–24.
Ross, D. (2000). Introduction: The Dennettian stance. In D. Ross, A. Brook, & D. Thomson (Eds.), *Dennetts philosophy: A comprehensive assessment* (pp. 1–26). Cambridge, MA: MIT Press.
Ryle, G. (1969). *On thinking*. Totowa: Blackwell.
Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.
Shannon, C. (1950a). A chess-playing machine. *Scientific American, 182*(2), 48–51.
Shannon, C. (1950b). Programming a computer for playing chess. *Philosophical Magazine 7th Series, 41*(314), 256–275.
Shepard, R. (1966). Learning and recall as organization and search. *Journal of Verbal Learning and Verbal Behavior, 5*, 201–204.
Sherrington, C. (1947). *The integrative action of the nervous system*. Cambridge, UK: Cambridge University Press.
Skidelsky, L. (2006). Personal-subpersonal: The problems of inter-level relations. *Protosociology. Special Issue: Compositionality. Concepts and Representations II: New Problems in Cognitive Science, 22*, 120–139.
Skinner, B. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.
Skinner, B. (1957). *Verbal behavior*. New York: Appleton.
Smart, J. (1959). Sensations and brain processes. *Philosophical Review, 68*, 141–156.
Smart, J. (1970). Review to *content and consciousness. Mind, 79*, 616–623.
Spencer, H. (1855). *The principles of psychology*. London: Longman, Brown, Green, and Longmans.
Taylor, C. (1964). *The explanation of behavior*. London: Routledge and Kegan Paul.
Thorndike, E. (1905). *The elements of psychology*. New York: Seiler.
Thorndike, E. (1911). *Animal intelligence: Experimental studies*. New York: Macmillan.
Tononi, G. (2012). *PHI: A voyage from the brain to the soul*. New York: Pantheon Books.
Turing, A. (1948). *Intelligent machinery*. National physical laboratory report. Reprinted: Meltzer, B., & Michie, D., (1969). *Machine intelligence* (pp. 3–23). Edinburgh: Edinburgh University Press.
Turing, A. (1950). Computing machinery and intelligence. *Mind, 50*, 433–460.
Tye, M. (2009). *Consciousness revisited. Materialism without phenomenal concepts*. Cambridge, MA: MIT Press.
Uttal, W. (2003). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.
Vogt, C. (1847). *Physiologische Briefe für Gebildete aller Stände*. Stuttgart: Cotta.

von Bertalanffy, L. (1968). *General system theory: Foundations. Development. Applications*. New York: George Braziller.
von Foerster, H. (1974). *Cybernetics of cybernetics*. Urbana: University of Illinois.
von Helmholtz, H. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Braunschweig: F. Vieweg.
von Helmholtz, H. (1867). Handbuch der physiologischen Optik. In G. Karsten (Ed.), *Allgemeinen Encyclopädie der Physik IX*. Leipzig: Leopold Voss.
von Neumann, J. (1945). First draft of a report on the Edvac. reprinted 1993. *IEEE Annals of the History of Computing*, *15*(4), 27–75.
von Neumann, J. (1951). The general and logical theory of automata. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior: The hixon symposium* (pp. 1–31). New York: Wiley.
von Neumann, J. (1955). *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer.
von Neumann, J. (1958). *The computer and the brain*. New Haven: Yale University Press.
Walter, G. (1950a). An electromechanical animal. *Dialectica, 4*, 42–49.
Walter, G. (1950b). An imitation of life. *Scientific American, 182*(5), 42–45.
Walter, G. (1951). A machine that learns. *Scientific American, 185*(2), 60–63.
Walter, G. (1953). *The living brain*. London: Duckworth.
Weber, E. (1851). *Die Lehre vom Tastsinn und Gemeingefühl – auf Versuche gegründet*. Braunschweig: Verlag Friedrich Vieweg und Sohn.
Wiener, N. (1948). *Cybernetics. Or control and communication in the animal and machine*. Cambridge, MA: The Technology Press.
Wundt, W. (1871). *Grundzüge der physiologischen Psychologie* (2 Vols.). Leipzig: Engelmann.
Yerkes, R. (1907). *The dancing mouse. A study in animal behavior*. New York: Macmillan Company.
Yerkes, R. (1911). *Introduction to psychology*. New York: Holt.
Young, Z. (1964). *A model of the brain*. Oxford: Clarendon.
Young, Z. (1965). The organization of a memory system. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 163*(992), 285–320.

Chapter 2
A Most Rare Achievement: Dennett's Scientific Discovery in *Content and Consciousness*

Don Ross

Abstract The chapter re-visits Daniel Dennett's first book, *Content and Consciousness* (1969), after four decades of developments in cognitive science and related disciplines. It first argues that in that book Dennett reported a scientifically significant discovery about what minds are. This initially seems implausible, because at first sight *C&C* presents as an exercise in pure philosophical analysis of everyday discourse about the mental, and that is a profoundly unlikely method for achieving scientific progress. However, a reading of the text and its context is proposed that explains this apparent miracle. The pure philosophical analysis indulged in *C&C* merely serves to blunt the force of previous philosophy, for the benefit of those who might find it persuasive. Thereafter, the positive discovery for which Dennett deserves credit comes as a specimen of the only kind of contribution to objective knowledge to which any philosopher (qua philosopher) can aspire: unification of empirical findings. The chapter then argues that because *C&C* does not try to integrate its unifying suggestions with any considerations from physics, it fails to offer a satisfying metaphysical account of the mental, even though most philosophical readers would see that as having been one of its central ambitions. Two decades after he wrote *C&C*, however, Dennett showed how to begin to close that gap. The chapter closes with reflections on differences between Dennett's view of the potential contribution of philosophy to science and the view of James Ladyman and Don Ross.

D. Ross (✉)
University of Waikato, Hamilton, New Zealand

University of Cape Town, Cape Town, South Africa

Georgia State University, Atlanta, GA, USA
e-mail: don.ross931@gmail.com; don.ross@uct.ac.za

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_2

2.1 Introduction

Do philosophers ever discover novel general truths about the world?[1] To judge from regular acerbic comments by scientists, and no small number of philosophers themselves, one would think not. This should not be considered surprising. The philosopher who sets out to produce original knowledge or insight typically uses the resources of his or her own brain and the network of informational connections in which it is embedded. Though no one else's brain is identically so embedded, most other philosophers who have explored the same conceptual terrain are very similarly situated – most knowledge in any academic discipline is, by institutional design, massively redundant. And though a human brain is an enormous processor as information storage-and-manipulation devices go, without profoundly novel input it will seldom generate profoundly novel output, and it cannot much improve the frequency of this through mere effort. Of course philosophers read a steady stream of new work by other philosophers. As Ladyman and Ross (2007), among others, argue, however, such dialectics tend to degenerate into the intellectual equivalent of stagnant ponds. Replacing the real fuel of empirical discoveries by the merely apparent energy of other philosophers' ruminations, debates become largely semantic exercises, at their worst implicit legislation of language by a group of people whose orders have no prospect of being followed.

The foregoing complaint about institutional philosophy as a whole does not imply that individual philosophers who engage the worlds of science, politics, law and art outside the confines of their discipline never arrive at arresting associations of ideas that influence wider cultures. Among twentieth-century philosophers, consider Russell or Rawls, for example. Based mainly on his later books, aimed at broad audiences, on consciousness, on Darwin's cultural impact, and on religion, many would recognize Dennett as another philosopher who has broken out of the airless cloister to achieve broad cultural influence. But acknowledging this falls short of agreeing that Dennett did something that quite a lot of scientists claim *never* happens: discover a non-mathematical fact of permanent scientific importance just by thinking.[2]

I will argue that although scientific discovery achieved by reflection on the structure of an idea is extremely infrequent and unlikely, Dennett's example shows us that it is not impossible. Dennett noticed something that, before it was spotted, seriously retarded progress in the behavioral sciences (including the social sciences). This proposition, since its recognition by Dennett, has been sewn into the basic fabric of these sciences and is unlikely to ever be withdrawn from among the set of

[1] The question here is not whether they make discoveries about neglected textual elements of the history of the philosophical corpus itself. I once heard the distinction I am making enunciated as follows by a distinguished philosopher – sadly, I forget which one. "In our Department," he said, "we have bullshit; and then we also have the history of bullshit."

[2] Most scientists agree that Einstein made an important scientific contribution with his famous thought experiment about an observation platform on a moving comet. However, few would argue that that contribution was important *by and in itself*; it was a minor aspect of a much wider achievement brought about by more conventional scientific activity.

acknowledged truths. Its discovery is announced in Dennett's first book, *Content and Consciousness* (1969) (hereafter, *C&C*). Unlike most of Dennett's later work, which synthesizes empirical findings from cognitive science to repair confusions fostered by non-scientists, *C&C* appears to be a straight exercise in the standard style of analytic philosophy. That is, it uses comparative consideration of various popular and scientific locutions, taken as expressions of judgments about reality, to try to tease out a consistent concept of the mental. On this basis certain common assumptions about the nature of minds are pronounced incoherent and an alternative, novel, set of assumptions is proposed that better accommodates the balance of well motivated judgments, including scientific ones. This sort of activity is almost never of scientific significance. A main reason for this is that natural language is shaped by a multitude of pressures and functions, of which expressing general judgments held to be contingent on objective evidence – in other words, proto-scientific generalizations – is at best a minor and transient one. So finding an instance in which philosophy based on semantic analysis produced a discovery of scientific importance is akin to stumbling across the Kimberly diamond.

What Dennett discovers in *C&C* is built upon a negative proposition: if there are minds, they are not among a wider class of entities that can be identified with material objects as such objects were understood by physics before quantum mechanics. To fill the resulting ontological void, the positive discovery is then the following: a mind is a narrated compression of patterns in the dispositions, embodied in differential neural network weights, conditioned by an individual's history of encounters with the environment while she tries to keep herself relatively secure, calm and occupied.[3] The summary is mainly supplied by the person herself, in response to

[3] An editorial reviewer objects that it "sounds like a category mistake to say that, according to Dennett, minds can exist *as* representations ("narrated compressions") of dispositions". The basis of the objection seems to be metaphysical, or perhaps semantic: the class of first-order existents and the class of representations are asserted, presumably on the basis of a priori analytic insight, to be disjoint. However, this denies what I have previously argued – and Dennett has agreed – to be one of his central philosophical claims, to whit, that (1) the mind is a virtual object (a 'Joycean machine') implemented by the brain's interactions with its environment, and (2) that being a virtual existent is a way of being a *real* existent, not a way of being primed for ontological reduction or elimination (See Sect. 2.3 below.) The kind of virtual object that the mind is is: a narrative. The key Dennett texts in this regard are Dennett (1991a, b). Then see Ross (2000a), with Dennett's (2000, pp. 356–362) reply; and Ladyman and Ross (2007, Chap. 4). For another supporting text, see Zawidzki (2007). Then the present chapter presents the case for thinking that Dennett had already arrived at the substance of this position, even if it wasn't yet fully explicit, in *Content and Consciousness*. I might add that since, as a naturalist, I deny that there are such things as reliable a priori analytic insights, I also deny that are, strictly speaking, such things as category mistakes (though of course there are uses of words so idiosyncratic that others won't understand them without explanation). The reviewer also objects that when Dennett's view of mind is "ontologized", it becomes a version of functionalism, in which case its originality as asserted here is undermined by the fact that Sellars and Putnam were functionalists before Dennett came along. Of course I am not claiming here that Dennett owes no intellectual debts, or that he is not a *kind of* functionalist. However, I do not think that a reader of Sellars would be likely to extract the full conception of mind I here credit to Dennett unless she had been led to go looking for it by prior acquaintance with Dennett. As for Putnam, he was an internalist about intensional content though he was an externalist about meaning more generally; and this is a crucial difference between his position and Dennett's, as commented upon in the main text below.

probes from others who must interact cooperatively and competitively with her and who thus have an interest in predicting what she will do. But self-observation is no more infallible than other observation, so sometimes an adult's construction of her own mind is corrected by others. The construction project is initiated by the infant person's immediate caregivers, though their role as principal mental engineers is supplanted by peers after late childhood.

This view of the mind is now the consensual ontological basis for several thriving enterprises in applied science and technology, specifically developmental psychology, social and cognitive neuroscience, and the efforts to build autonomous robots that can work for and with people. It also helps us make sense of such social sciences as economics and sociology, because it allows us to understand how behavior can be responsive to incentives – that is, broadly rational – without being produced by deliberate individual ratiocination. Dennett's understanding of minds as virtual constructs, built in service of and then maintained and stabilized by social coordination, explains how choices can be distributed in populations, without being internally computed by most or even any individuals, which in turn explains why economics and sociology do not reduce to individual psychology and neuroscience. The contents of an individual mind are, as Dennett (1991a and elsewhere) puts it, built through dynamic social equilibration among the narratives of multiple authors, even if the individual to whom the mind in question is attributed is the primary author. Thus facts about an individual's intentional profile, including her choices, are partly functions of irreducible social facts and processes (See Ross 2005 for detailed exposition).

The above paragraphs are written in light of the mighty wave of new knowledge in the behavioral and social sciences that has accumulated in the 45 years since *C&C*. The main business of the present essay is to identify more precisely how much of this can be said to be anticipated in *C&C* itself. I concentrate here only on the parts of *C&C* – Part 1, and Chap. VIII–X – that concern the general character of minds. I think that Dennett was additionally the first thinker to achieve a reasonably comprehensive grip on the nature of human self-awareness; but I do not think this was evident until Dennett (1991a). In any event, however, consciousness is not my topic here.

In the following essay I will argue that the achievement of *C&C* in seeming to accomplish a major advance in empirical understanding on the basis of an exercise in analytic philosophy is less miraculous than it at first appears. The pure philosophical analysis in *which C&C* engages merely serves to blunt the force of previous philosophy, for the benefit of those who might find it persuasive. Thereafter, the positive discovery for which Dennett deserves credit comes as a specimen of the only kind of contribution to objective knowledge to which any philosopher (qua philosopher) can aspire: unification of empirical findings. I will also argue that because *C&C* does not try to integrate its unifying suggestions with any considerations from physics, it fails to offer a satisfying *metaphysical* account of the mental, even though most philosophical readers would see that as having been one of its central ambitions. Two decades after he wrote *C&C*, however, Dennett showed how to begin to close that gap, as I will explain.

2.2 C&C in the History of Cognitive Science

Among the more remarkable features of *C&C* is the extent to which it was alert to the then newest sources of scientific excitement without being critically overwhelmed by them. It was among the earliest work in the philosophy of mind that fully incorporated the cognitivist challenge to behaviorism that quickly gathered steam following Chomsky's (1959) famous criticism of Skinner. The language used to express this point in *C&C* is now dated, but the substantive content is not. "Centralist" accounts of mind, meaning accounts that traffic in hypotheses about internal representations, are required in place of "peripheralist" models that refuse to posit any internal mediators between environmental influences and overt behavior except conditioned mechanisms (*C&C*, p. 43). Abstraction and generalization, required for systematic response to complex patterns and ideas, are the very essence and point of mindfulness. Dogmatic behaviorism that restricts attention by methodological fiat to stimulus-bound perception and action thus fundamentally cuts itself off from the possibility of modeling or explaining mental phenomena.

We know, as the author of *C&C* could not, that what followed in the most dominant cognitive science and philosophy of mind of the 1970s was an equally self-defeating rush to the opposite monistic pole, where the importance of the storage of some information in the external environment and in motor routines was forgotten. Internal representation, the formerly neglected necessary condition for mindfulness, became widely regarded as *sufficient* for it. This reached in apogee in Fodor's (1980) plea for "methodological solipsism" in cognitive science, the idea that one should study internal representations without regard for the environmental influences to which they are adaptations, and denying that conditions of behavioral application are relevant to semantic interpretations of mental content. This was in order that the semantics of representations could be modeled as a strict function of their 'syntax'. The point of *that* was that, according to Fodor, and many like-minded theorists such as Pylyshyn (1984), only syntactical differences could make causal-mechanical differences to the production of behavior.

C&C, notwithstanding its pioneering cognitivism, is in no way ancestor to this anti-behaviorism run riot. It acknowledges that the evolutionary basis of mind is the fitness value of being able to learn, and that the basic form of learning in people as in other animals is conditioning (*C&C*, pp. 62–63). But it emphasizes that conditioning works through the sculpting of weights in networks of neurons. Between stimulus and response – not *instead of* stimulus and response – lies computation.

The brief account of neural learning given in two pages of *C&C* (pp. 55–56) could still be taught as an introduction to this topic in 2012. This is *not* because it is so abstract that it was bound to accord with whatever neuroscientists subsequently discovered. The best of the cognitive neuroscience read by the young Dennett – especially Hebb (1949)[4] – was exemplary science that resisted the impulse to

[4] Hebb is not cited in *C&C*. But Dennett (personal correspondence) affirms that he had read Hebb, along with McCulloch, Ashby, Grey Walter, Michael Arbib and J.Z. Young.

speculate beyond what was genuinely empirically established, and so sketched a general model of the learning capacities of networked synapses that has since been massively enriched in terms of its mathematical characterization (Sutton and Bartow 1998) and its underlying basis in neurochemistry (Reynolds et al. 2001; Seung 2003), but not fundamentally revised.[5] The most salient point in light of later philosophical arguments is that Dennett recognized that Hebbian neural computation is a form of conditioned learning, not an alternative to it. This was neglected by most cognitive scientists, and vehemently criticized by leading philosophers of mind,[6] especially during the early stages of the resurgence of connectionist models of cognition in the late 1980s. A main reason that *C&C* holds up much better than other philosophy inspired by the first generation of cognitive science is that it did not fall into or fan the over-reaction against behaviorism.[7]

Dennett did not resist hyper-internalism about intentional content only after philosophers infatuated with sweeping general theories pushed them far past any basis of support in empirical psychology. In *C&C* he sees the language of thought (LOT) hypothesis (Fodor 1975) coming and pronounces it implausible (*C&C*, p. 87). At the very end of the book he announces that "Thoughts […] are not only not to be identified with physical processes in the brain, but also not to be identified with logical or functional states or events in an Intentional system (physically realized in the nervous system of a body)" (*C&C*, p. 189). Here we find an unqualified rejection of computational functionalism, roughly a decade before it briefly flourished as the dominant theory of the ontology of the mental among philosophers and the many scientists in AI who concurred with them. In 1988 the originator of that account, Putnam (1988), renounced it, and it thereafter slowly faded from prominence in philosophers' discussions. Dennett's skepticism about it two decades earlier does not seem to be widely acknowledged, perhaps because his later defense of a much less restrictive form of functionalism against Searle (e.g. Dennett 1980) gained much attention.

According to *C&C*, then, mental representations are not to be identified with literal brain occurrences of tokens isomorphic in 'logical form' to linguistically structured content, but are instead constructed on the basis of behavioral observations and probes designed to reveal aspects of supporting neural computations. Such probes in the case of adult human subjects could consist in straightforward verbal questions. But of course psychologists have long complemented these with other kinds of probes, especially asking subjects to perform non-linguistic tasks in varying conditions under experimental control. This opens doors to the study of the

[5] An even earlier general, non-technical account of neural learning sketched for philosophical purposes that also – in this case explicitly – owes its accuracy to reliance on Hebb is Hayek (1952); see Ross (2011).

[6] See, for example, Fodor and Pylyshyn (1988). Philosophers, I speculate, had a professional interest in seeing mental modeling turn out to be an exercise in applied formal logic. Hence their widespread advocacy of so-called classical artificial intelligence.

[7] Dennett cannot plead completely innocent here, however. While understanding why the title seemed irresistible, I wish that he had not later written a paper called 'Skinner skinned' (Dennett 1978).

emergence of implicit abstract representation in two groups of agents that cannot be asked to report their thoughts using linguistic paraphrases: non-human animals and young children. And this, after Dennett in later work replaced the philosophical analysis of *C&C* with more direct methods of presentation and argument that were amenable to scientists, eventually provided the channel by which the discovery reported in *C&C* actively transformed areas of scientific inquiry. False belief protocols, in which cues for alternative inferences are delivered to experimental and control groups of children and animals, became the standard method for constructing models, framed as intentional descriptions, of the coupling of infant and non-human information-processing capacities with environmental contingencies. Pioneers of these methods explicitly cite their inspiration by Dennett (Griffin and Baron-Cohen 2002; Seyfarth and Cheney 2002). The Dennettian ontology of minds as equivalent to such couplings thus became the basis for the foundational methodology of cognitive ethology and cognitive developmental psychology. Prominent among the areas of inquiry for many of these scientists have been the idiosyncratic theories of mind revealed by the behavior of monkeys, apes, corvids, and others when they interact strategically, and the maturation of such theories toward standard adult models in human children. These experimental designs were unlikely to have occurred to anyone who imagined that minds were identical to brains, or that they required internalization of the full structural articulation of human language.

A number of theorists besides Dennett recognized independently that both environmental pressures and internal computational organization are necessary conditions, with neither being sufficient, for the emergence of mindfulness.[8] However, it was Dennett who first gave wide currency to the idea[9] that what mind *is* is a model of the patterns that systematically *link* these two kinds of structure.[10] Mind is a unique kind of model in not being, like most models, merely an aid to explanation and prediction by curious scientists or ambitious engineers. It is instead a model that arises naturally and is embedded in phenomena that date back to the origins of (at least) modern humans. The very point of internal computational organization is to be coupled with environmental patterns, and in a highly social species with a massive, and massively plastic, brain. The coupling in question must furthermore be an object of joint coordinated attention by people who live and work together. This is

[8] A clear instance is Bruner et al. (1956), founders of the 'New Look' that brought active internal representations back into experimental psychology. Bruner (1990) later expounded a narrative model of the integrated conscious self, around the same time that Dennett did.

[9] There is a good case to be made that Vygotsky (1934/1986) was the first to expound this idea in press. But Vygotsky's influence was mainly confined to a literature in the psychology of education that stayed isolated from broader philosophical debates until the late twentieth century. Dennett (personal communication) says that he had not yet heard of Vygotsky when he wrote *C&C*.

[10] An editorial reviewer again objects (see note 3) to my treating the mind, on Dennett's view, as a kind of existent and also as a kind of representation (a "model"). Again, however: according to Dennett minds are narrative structures. The narratives in question link patterns of brain responses to patterns of environmental contingencies. Since many such relationships among patterns are *not* represented in the narratives in question, or are emphasized or de-emphasized relative to others, the narratives, and thus the minds they constitute, are models.

why none of us can evade the pressures to build and maintain viable models of ourselves. Dennett was not the first to see that mind is a kind of process. He was, I suggest, the first to recognize that the kind of process it is is a *descriptive* process, as I will explain.

2.3 Instrumentalism and Realism

In the Western philosophical tradition, to say that something is a kind of description rather than a kind of thing (or rather than a kind of description-independent process) is a way of saying that it is not real. And indeed throughout his career, from the publication of *C&C*, Dennett has been associated with instrumentalism about the mind, that is, with the idea that the mind is 'merely' a construction and in that sense not part of the genuine furniture of the world.

Dennett at one time accepted the instrumentalist label, but later retracted this concession (Dennett 1993, p. 210). Gradually he developed a defense of the idea that being a socially constructed and maintained narrative description is a way of being real rather than a way of being 'less than' or only 'semi' real. These efforts culminated in his well known paper "Real patterns" (Dennett 1991b), which argues that minds are virtual objects, and in this status resemble most of the abstract existents about which the social and behavioral sciences construct and test hypotheses and models. Ross (2000a) and Ladyman and Ross (2007, Chap. 4) argue that Dennett's (1991b) account still falls short of proposing a fully fledged realism about the virtual mind. However, they maintain that Dennett provides himself, and us, with the resources for such a realism when he points out that an observer who failed to represent people's minds would thereby miss information about the world, even if she had a complete record of their brains and their external environments, because she would then not have an account of which aspects of brains and environments (including other brains) were systematically functionally coupled.[11] But to say that failure to identify or describe a pattern implies incompleteness of objective information – that is, information that would have to be included in a complete science – is the only legitimate litmus test for judging something to be real. Thus, one should be a realist about minds, about the world's US Dollar savings, about rates of monetary inflation, and about many other virtual entities. But, to take an example from Ross (2000a), one should not be a realist about the entity 'my left nostril, the Namibian government and Miles Davis's last solo' because this nominal entity is redundant – tracking it carries no information not captured by tracking the three components to

[11] Note that in this reasoning Dennett captures the truth in Jackson's (1982) recognition that a purely neuroscientific account of a person would miss some facts about her, without falling into Jackson's error of supposing that the fact in question must be *only* about her rather than about her relationship to the environment. Jackson's error has had the most retrograde of possible consequences, turning in the hands of Chalmers (1996) into a revival of dualism.

which it reduces. If minds reduced to brains or to behavior they too would be redundant and not real. But they do not so reduce.

In light of this subsequent history, one is not surprised to find that *C&C* describes the mind in ways that suggest instrumentalism. For example, Dennett refers to intentional description as "a heuristic overlay" on extensional accounts of the variables that are used to model the neural causes of behavior (*C&C*, p. 80). A few pages later he says that "[…] although such systems [brains as engines of purposive behavior] are ultimately amenable to an extensional theory of their operations, their outward manifestations are such that they can be *intelligibly* described at this time, within our present conceptual scheme, only in the Intentional mode" (op. cit., p. 89). This attitude is not far from that of eliminativists such as Churchland (1979, 1981), according to whom minds, construed as intentional descriptions of the networks of internal causation of behavioral regularities, are crutches we use for now, while we wait for neuroscience to furnish the true model of the real mechanisms in its own conceptual terms that will displace – not reduce – the intentional domain. Dennett's formulations in *C&C* are thus not always consistent with his recognition, which later became more salient, that "outward manifestations" are partly (but only partly) *constitutive* of the mental.

It is instructive that of all the articulations of his theory of the mind that Dennett has produced over the course of his career, the one most strongly based in traditional philosophical analysis and argumentation got the science of the story right in all its essentials – as judged against both Dennett's later opinions and what has been implicitly endorsed by later scientific practice – while providing an unstable and unsatisfactory account of the metaphysics. In the remainder of the discussion I will pursue some morals from this.

2.4 Useful Philosophy and the Unification of Sciences

It is a common theme in recent commentary on science that the heroic phase of enlightenment history is over. (See Humphreys 2004 for a superior instance of this theme.). The earliest breakthroughs in each discipline were achieved partly through fundamental conceptual reorganizations leveraged by invention of new mathematical technologies. Thus one can, without too much distortion, write the early history of science as intellectual biographies of great individual thinkers who established the foundations of the disciplines: Galileo, Newton, Lavoisier, Darwin, Einstein, von Neumann. But by the middle of the twentieth century the struggle to break free of folk ontologies and conceptual "cul-de-sacs" had been won everywhere except in the social sciences. In physics, chemistry, and biology Sellars's 'manifest image' simply became irrelevant, rather than being the basis of a worldview that needed to be effortfully transcended in forging new technical alternatives. The training of an expert in any of these disciplines now involves absorption in a specialized framework literally describable only by mathematics, which then anchors the semantic structure of a closed natural-language argot for each discipline. It is an important

mistake to think that in such cases the mathematics more precisely represents the content of ideas originally cast in natural language; the truth is rather that the jargon comes later and refers to operations performed with the mathematics. This is why contemporary scientific language often sounds like a kind of slang. If scientific language *were* a refinement of everyday language, one would expect there to be a useful role for those with special expertise in jimmying everyday conceptual networks built for practical purposes into more orderly edifices that respect logic and at least accommodate science even if they cannot be used to accurately express its latest and most exotic findings. However, once we recognize that scientific language is instead mainly a device for pointing to edifices of achieved mathematics and experimental practice, it becomes unsurprising that scientists do not usually find philosophers adding value, and indeed typically find their pontifications naïve.

Humphreys (2004) takes this story one step further than most other accounts. At the cutting edges of the most successful sciences, he argues, massive new computational capacity is pushing an ever larger share of scientific reasoning into the manipulation of statistical inferences that human brains cannot follow in detail. In this context, scientists must increasingly surrender the ambition even to enjoy arcane conceptual systems that make them feel at home in their specialized domains, and learn to relax in vertiginous intellectual environments where operationally meaningful 'ideas' are mere trained hunches about what statistics packages are doing.

This is a persuasive picture of what is happening in most branches of physics. It is also happening in my own discipline of economics, which poses special problems of entry and re-entry because the discipline is held responsible for offering policy advice, and recipients of such advice typically want at least a sketch of underlying reasoning that they can understand. As for psychology, it is running behind economics in terms of the range of statistical techniques with which practitioners are expected to be practically familiar. However, the conceptual strangeness emerging from psychometrics is deeper than that handled by the econometrician because psychologists' much greater rate of traffic in latent constructs adds an entire dimension of complexity to statistical inference modeling that economists and econometricians try to avoid. That is, psychologists, but not economists (though see Andersen et al. 2014), must parse their catalogue of constructs into different kinds for purposes of statistical treatment. Psychologists also face greater resistance than economists when new evidence leads them to draw conceptual distinctions that folk ontology does not recognize.

What does all of this have to do with the singular achievement that, I have argued, can be found in the pages of *C&C*? Dennett (2013) has recently articulated a conception of the current and continuing value of philosophy that might seem at first sight to be beautifully exemplified by *C&C*. This conception, of the philosopher as reconciler of the manifest and scientific images, is in tension with an alternative understanding of the point of philosophy, as the discipline for constructing a unified scientific image that supplants the manifest image, defended by Ladyman and Ross (2007) (and elaborated further in Ross (2013) and Ladyman and Ross (2013)). Their picture of the philosopher's role at its best might *also* be thought to be nicely on display in *C&C*. This might lead one to suggest that the example of *C&C* points the

way to dissolving the conflict between the two visions of philosophy's future. However, I will give reasons for rejecting this suggestion. Like many great texts in the history of science (Kuhn 1957), *C&C* is Janus-faced, simultaneously a late hurrah for a tradition that was expiring and at the same time a pioneering rough prototype of philosophy's future.

2.5 Dennett and the Role of Philosophy

In his recent thoughts on the relationship between science and philosophy, Dennett (2013) says that "[…] a large part of philosophy's task […] consists in negotiating the traffic back and forth between the manifest and scientific images" (p. 99). This sort of mediation between conceptual spaces has been a preoccupation of Dennett throughout his career, as has been noted before (e.g. Ross 2000b). However, within his conception of this activity over the years is to be found some ambiguity. In his work beyond *C&C* he emphasized tolerance for alternative 'stances' on phenomena of study. These stances – physical, design and intentional – refer to different ways in which real domains can be conceptually parsed for varying practical purposes. If we were to read him as meaning that some of these practical purposes are scientific while others (e.g., technology development or policy guidance) are not, this would support the common instrumentalist reading of Dennett's intentional and design stances. However, this would be a mis-reading of his mature stance on stances, at least following his explicit renunciation of instrumentalism about the mental. The different purposes supported by each of the physical, design, and intentional stances on mind and behavior involve both prediction and explanation, and Dennett takes all three to be relevant to realistically interpreted science. Sherlock Holmes paints a subtle portrait of the murderer from the intentional stance and in so doing both explains why the murderer did it and helps us predict the specific form of menace she presents to particular types of others. Then when a victim's grieving relative cries "But why?" and means it existentially, the evolutionary psychologist can explain from the design stance why there is always a non-zero frequency of such murderers in every population. Meanwhile the neuropsychologist assumes the physical stance and explains that our criminal responds to status competition and other forms of conflict *murderously* – rather than, say, with an unusually aggressive performance on the tennis court – because some key genes for orbitofrontal GABA signaling are underexpressed in her relevant cortical RNA pathways. For a complete *objective* account of the murderer all three stances have irreplaceable work to do.

On this picture, we do not have a folk stance that lives in uneasy coexistence with a scientific stance; the intentional stance is a necessary part of the behavioral and social *sciences*. In consequence, unification of these sciences with their more mechanistically oriented neighbors such as biochemistry requires conceptual mediation within the scientific image. Because the phenomena must be consistently captured across all stances – inter-stance contradiction must be avoided – none of the stances should be expected to extend folk ontologies without substantial revision

(Ross 1994). Consider, for example, Dennett's view of consciousness. Folk psychology, regarding it as a Cartesian theatre, does not anticipate Dennett's view of it from the design stance as a virtual machine installed by cultural pressure on the neural network hardware of the brain, nor his intentional stance characterization of its contents as determinate only *after* probing by oneself or another with a demand for an explicit account using a natural public language.

Mediating among stances to preserve ontological unity within the scientific worldview[12] is a plausible role for a scientifically well informed philosopher. According to Ladyman and Ross (2007), this is the *only* way in which the philosopher qua philosopher can make a professional contribution to objective knowledge. The need for diverse stances on data that fundamental physics treats as the same data arises, according to Ladyman and Ross, from the scale relativity of ontology, which in turn arises from the fact that stochastic processes typically give rise to emergent regularities only when systems become sufficiently large and complex for incomplete informational redundancy to have statistically estimable effects. An especially controversial aspect of Ladyman and Ross's view of the scientific basis for sound metaphysics (Melnyk 2013) is that, according to them, unification only constitutes metaphysics when one of the stances entering into the unification is drawn from fundamental physics; unifications between special sciences are simply theoretical innovations in those special sciences. I will return to this issue later. In the meantime, I note only that Dennett's picture of the philosopher as the mediator among scientific stances is one of the core intellectual inputs to Ladyman and Ross's naturalistic metaphysics[13] and its associated philosophy of science; and that philosophy is as radical and unfriendly to conservative conceptual structures as any version of eliminativism.

[12] This is what Ladyman and Ross (2007), broadly following Kitcher (1981), mean by 'unification'. Unification usually means something stronger in the philosophy of science: specifically, Nagelian intertheoretic reduction. But Ladyman and Ross argue that this fails to capture the actual activity that scientists regard as unifying; and to define unification in the narrower way begs the question against these arguments.

[13] The project of Ladyman and Ross (2007) can be summarized as follows. First, they argue that analytic metaphysics based on intuitive folk ontologies rather than fundamental physics is highly unlikely to succeed in discovering any objective truth, both for epistemological reasons and because folk ontologies are in fact incompatible with the ontology of our actual fundamental physics, quantum theory. After defending a version of structural realism in philosophy of science against standard realist and constructive empiricist alternatives, Ladyman and Ross construct a naturalistic metaphysics based on the structures identified by quantum theory. The task of such a metaphysics is to provide a general account of objective reality by explaining how the proliferation of ontologies reflected in the special sciences can all be true of a single world. According to this metaphysic, reality is irreducibly stochastic but has enough structure to support true statistical generalizations. These generalizations describe 'real patterns' in a sense that refines the original idea of Dennett's (1991b). The patterns successfully studied by special sciences, many of which are characterizable in the structures of Sellars's 'manifest image,' are real; but the more pervasive real patterns that unify them, those identified by fundamental physics, can be represented only in mathematics, not using the subjects and predicates of natural language. Thus metaphysics naturalized by reference to physics cannot be expressed in such language. For commentaries on the project, including Dennett's own, see Ross et al. (2013).

On the other hand, Dennett has also often suggested, and particularly explicitly in Dennett (2013), that the philosopher's job description involves mediation between conceptual frameworks in a wider sense. In particular, mediation between folk conceptual spaces and scientific ones – that is, between the manifest and scientific images – is regarded as activity similar to what is involved in brokering among scientific stances, and as calling upon roughly the same philosophical skill set.

I will not here take issue with the suggestion that the two kinds of mediation depend on similar aptitudes. However, quite different issues arise when one queries the point and value of reconciling the manifest and scientific images. It is surely not controversial that scientific findings must often be translated into everyday linguistic and conceptual terms so that policy makers, judges, juries, doctors, patients, soldiers, and many others can make decently informed decisions. And of course a public that pays for most basic science through its taxes appreciates efforts to make the resulting discoveries partly comprehensible to non-specialists. But these are not the motivations that Dennett has tended to emphasize. He argues instead that critical comparison of folk and scientific conceptual spaces is the core of an undersung part of science, cognitive anthropology. Dennett (2013) (see also Dennett 1991a, pp. 82–83) forcefully reminds the would-be cognitive anthropologist that folk ontologies are unreliable, and that the anthropologist must be careful to avoid "going native" by treating them as true. Thus efforts to "negotiate the traffic back and forth" between folk and scientific conceptual networks are not, on Dennett's view, made for the sake of enriching the science of the networks' objects of study (except literally – they might be important for keeping the funding taps open). Academic anthropology aside, they are for the sake of enriching general human experience. This aim, though always on his agenda (e.g. Dennett 1984), seems to have become increasingly central to Dennett in the later stages of his career.

Ladyman and Ross (2007) do not deny that explaining science to non-scientists is worthwhile activity. However, their view of this activity differs from Dennett's in two respects.

First, they argue that whatever services reconciliation of the manifest and scientific images might render for political and economic support of science, it tends to interfere with the epistemic progress of science. It has this effect because it encourages proliferation of analogies between scientific and folk ontologies, which invariably 'domesticate' the former in the sense of blunting their most radical implications for further conceptual revisions that in turn open roads to new experiments and new mathematical and statistical tools. Consider, for example, quantum mechanics. Efforts to make it comprehensible within familiar categories of being and logic have convinced almost everyone other than the purest experimental physicists that it requires an "interpretation", meaning some addition to the literal mathematics that "makes sense" of it. In consequence, almost all philosophers of physics and a number of theoretical physicists on their philosophical Sundays exert energy trying to choose between (e.g.) Bohmian realism about wavefunctions, Ghirardi, Rimini and Weber realism about spontaneous wavefunction collapse,[14] Everettian realism

[14] See Allori et al. (2008) for the ontological reading of Ghirardi, Rimini and Weber.

positing multiple branching universes to realize each trajectory consistent with every wavefunction, and so on. This has led to neglect of Bohr's version of the Copenhagen interpretation – which does not count as an "interpretation" in the first place according to most philosophers – as mere formalism without "physical content". Ladyman and Ross (2013) reject the intuitions underlying the demand for "physical content".[15] They arise, Ladyman and Ross argue, simply from the domesticating impulse. Associating failure to respect this impulse with "pure" formalism is profoundly unjustified in the context of the history of physics.[16] Among other relevant considerations, there is no such thing as purely formal statistics; all of the statistics of Bohrian QM[17] are derived from experimental data. All that is "philosophically wrong" with Bohrian QM is that the realist cannot say what she is a realist about using natural language. This does not imply that one must embrace instrumentalism. Ladyman and Ross (2013) further contend that respect for the urge to domesticate has interfered with (at least) the speed and frequency of exploration of open avenues in physics. In particular, they argue that it has complicated the integration of quantum field theory with other wings of quantum theory, and has caused entanglement to be widely perceived as a problem instead of as the key to dissolving pseudo-problems, especially particle/wave duality (Ladyman and Ross 2007, Chap. 3).

The second point of Ladyman's and Ross's disagreement with Dennett on mediation between folk and scientific ontologies is that the former deny that philosophers are the best placed people to do it. As Dennett (2013) himself notes, philosophers persistently confuse cognitive anthropology with metaphysics, and so imagine that when they ponder semantically sound but scientifically empty questions, such as how many atoms can be removed from a statue before it loses its identity, they are trying to discover an objective fact about some aspect of reality aside from their own intellectual biographies (Astonishingly, it is common for them to think that their reflections, informed by no real physics at all, *succeed* in discovering such facts!). The moral here seems obvious. If you want cognitive anthropology done you should ring up some anthropologists (For an example see Atran 1990). Philosophers, by contrast, seem unable to resist going native – or, to phrase their problem more accurately, *staying* native. Ladyman and Ross (2007, Chap. 1) catalogue some egregious cases of this from recent philosophers working in the much-honored tradition of the late David Lewis. A very common, but equally crude, version of it is to treat fundamental physical particles as if they are like hard little separable bricks out of which larger objects are somehow glued together, when the

[15] Ladyman and Ross do not mean by this that physics need not be testable by physical experiments. They mean to reject the basis on which philosophers currently tend to draw the distinction between physics and mathematics.

[16] d'Espagnat (2006) is a physicist who agrees with this view of Ladyman and Ross's.

[17] I am cautious in my labeling here, because there are versions of 'Copenhagen' QM that postdated Bohr, according to which wavefunction collapse is a consequence of measurement. This *should* be regarded as a last resort on philosophical grounds because it is idealism, and as such looks presently impossible to unify with other parts of physics or with science generally.

actual sub-molecular world does not resemble that picture in any interesting respect whatsoever. A more subtle inability to break with the natives is expressed in the mistake discussed above of thinking that Bohrian quantum theory does not count as interpreted.

2.6 C&C as Philosophy and as Science

One strategy that makes it almost impossible to escape from staying native is taking intuitions based on linguistic usage as data. This brings us back to *C&C*. The reader familiar with Dennett mainly through his later work will be surprised, on coming to *C&C* for the first time, to encounter a flurry of analysis of everyday phrases about the mental. The origins of core aspects of Dennett's thought in Ryle's Oxford tutorials is suddenly revealed as surprisingly literal. As I intimated at the beginning of the present essay, one could hardly imagine adopting a method that is less promising as a road to scientific discovery.

Attention to the larger structure of the text, however, reveals not the use of ordinary language analysis (henceforth OLA) for positive purposes, but a final deployment of it *against* barriers to science thrown up by its own previous uses. Dennett is concerned to refute the proposition that a "problem of intentionality" blocks any effort to integrate the domain of the mental into the network of natural causal processes. The blockage in question consists of the following dogma: parts of language can be "about" other parts of language; but objects and processes cannot be "about" anything.[18] Thus intentional phenomena appear to be metaphysically *sui generis*. Dennett then uses OLA to show that previous exercises of OLA by philosophers promoting this problem themselves rested on selective attention to language. In particular, they failed to note that most phrases that implicitly associate virtual referents with the ontologies of objects function like idioms: most logical form properties of terms that refer to objects will not inferentially carry over to terms that refer to virtual entities so far as the shared cultural intuitions of speakers are concerned. This holds not only for terms for virtual objects, such as minds, about which philosophers have raised deep metaphysical puzzles, but about terms for less portentous virtual objects, such as voices, about which they have not. The implication of this is wholly negative: the behavior of English speakers shows that they are not collectively committed, after all, to regarding minds as kinds of things. This undermines attributing implicit *substance* dualism to the folk just as much as it undermines attributing physicalism to them; they talk about minds not as if they are physical objects *or* states of immaterial souls, but as if they are not normal *objects*

[18] This dogma is persistent. It is resurrected without qualification in Rosenberg's (2011) new argument for the objective non-existence of any form of meaning, for the impossibility of naturalistic philosophy, and hence for the irrelevance in principle of all philosophy to objective knowledge. Rosenberg does not critically consider Dennett's account of intentionality. He will not have completed his case for über-eliminativism until he does so.

at all, but instead occupy a suo generis ontological category. *If* one thought that OLA was a reliable method of pursuing any form of inquiry other than cognitive anthropology, one could interpret this conclusion as evidence for a sophisticated kind of dualism such as Chalmers's, according to which the mental is ontologically mysterious (Someone indulging in such reasoning would need to either agree that voices are also ontologically mysterious, or deny that speakers take the literal existence of voices seriously while insisting that they do, and that we should, take the existence of minds seriously).

But no argument in *C&C* requires accepting the premise that OLA is a reliable method. Dennett's indulgence of it is used only to reject conclusions derived from prior uses of OLA. Perhaps folk usage shows no settled ontology of mind because folk usage is seldom metaphysically consistent in general; or perhaps it is because the folk have not yet assimilated the new findings from new sciences – particularly, post-behaviorist psychology and neural learning theory – that show how to dissolve the mystery. I think that Dennett has always been ambivalent about the extent to which the folk "know what they're talking about". The structure of argument in *C&C* usefully shows why he has not felt a need to resolve this ambivalence. Whether you think that folk usage is inconsistent because the folk are not interested in consistency or because the folk have not yet absorbed enough new science, the next thing you should do as a philosopher is exactly the same: turn to the new science to see how it helps you formulate a stance that *is* consistent. That is just what Dennett does in *C&C*. His discovery that the domain of the mental – of the intentional – is the domain of coupled relationships between internal representations and overt actions and perceptual responses does not depend on any conclusions from exercises in OLA. Indeed, the OLA disappears from the text once he begins to describe the relevant science. The positive discovery that I opened the present essay by celebrating rests entirely on generalizing findings of the first generation of what we would now call cognitive neuroscience and *unifying* them with a Darwinian account of the selection of functional behaviors.

The Ladyman and Ross picture as outlined in Sect. 2.5 views this as an exercise in theoretical psychology, not metaphysics, because the unifying activity is not generalized by drawing in any resources from fundamental physics. This, I suggest, helps to explain why despite hitting a perfect bull's eye as psychology, Dennett's underlying metaphysics of mind – tilting unsteadily between instrumentalism and realism – remained uncertain and inconsistent. He achieved a major advance in metaphysical clarity two decades later, with "Real patterns", precisely when he reflected on mental phenomena as a special case of virtual phenomena more generally, in the context of a body of theory that is close to fundamental physics, namely, the theory of informational complexity.

At this juncture the dialectic swells in complexity beyond what can be usefully dealt with here. Some physicists (see Zurek 1990) think that information theory is fundamental physics. Ladyman and Ross (2007) are critical of this view, on grounds that the part of physics that makes spectacularly accurate predictions of out-of-sample measurements does not imply or rely on the second law of thermodynamics,

which is required for linking information theory to the *rest* of physics.[19] But all that is at stake here, in the present context, is whether grounding a general theory of mind in information theory is sufficient in itself for full metaphysical integration of the mental. Since no one doubts that some theoretical relationship between information theory and fundamental quantum theory is on the scientific agenda, the claim that Dennett's discussion in "Real patterns" represents progress toward a stable metaphysics of mind is also not in dispute.

I have argued that Dennett's negative conclusion in *C&C*, that minds cannot be identified with any class of pre-QM (i.e., 'classical') material objects, is based on philosophical analysis. Unsurprisingly, then, that analysis cannot soundly be used to establish the scientific need for a contrast class of empirically real entities that *can* be identified with classical material objects. Ladyman and Ross (2007) argue that discovery of the applicability of quantum physics to every measurable feature of the universe now shows us that there are no material objects as these were understood by classical physicists, and as they are still approximately understood by almost all non-physicists. Of course this does not call the negative conclusion about minds into question. Indeed, recognition that minds, which are not classical material objects, do not stand over against a contrast class of real classical material objects flushes the last vestige of Platonic and Cartesian dualism from the scientific worldview – Cartesian dualists are wrong about *both* halves of their duality – and thereby can be held to satisfyingly *complete* the project launched in *C&C*.

2.7 Conclusion

I have argued that Dennett achieves a genuine scientific discovery in *C&C*. I have also argued that, appearances to the contrary, this discovery was not implausibly realized through philosophical analysis. It was achieved by unifying several then recent empirical discoveries. The unification in question was incomplete and so fails to yield a satisfactory metaphysics of the mental; but Dennett later made some important progress in that direction when he appealed to information theory to relate mental phenomena to physical phenomena *as the latter are understood by physicists, rather than by the folk.*

This interpretation of *C&C* thus does not offer support to a conception of philosophy that views it as making potential contributions to objective knowledge by mediating between the scientific and manifest images. On that project, Dennett

[19] If the Everett interpretation of quantum mechanics is correct, then an alternative route to the fundamentality of information theory would be available. Quantum information theorists tend to be advocates of Everettian interpretation; see Deutsch (2010, 2011) and critical discussion in Ladyman and Ross (2013). Rosenberg (2011) bases what I referred to above as über-eliminativism (nothing objectively exists except fundamental physical particles) on conjoining the universality of the second law with a version of reductionism that seems to require ignoring entanglement; Deutsch (2011) would reject the second conjunct.

(2013) poses the following rhetorical questions: "Scientific utility, as Quine never tired of reminding us, is as good a touchstone of reality as any, but why shouldn't utility *within the manifest image* count as well? Is there anything dangerously relativistic in acknowledging that the two images may have their own 'best' ontologies, which cannot be put into graceful registration with each other?" (p. 105). These questions are ambiguous. Utility within the manifest image clearly counts for a great deal when one is trying to coordinate the actions and ambitions of people whose thoughts are wholly or partly structured by that image. In most aspects of social life, that describes all of us. In the quotation, Dennett wisely relates "reality" to "*scientific* utility". The manifest image is ultimately a *barrier* to scientific utility, even if we must frequently rely on it while staying practically afloat in our Neurathian boats, because it largely *misdescribes* the world. In particular, it dangerously misdescribes the *social* world as a struggle of individual wills and moral convictions, distracting people from the complex interplay of demographic and technological changes at the population level that are of much greater causal importance.

It is, however, possible to take this important point too far (as Rosenberg 2011 does), and deny that intentionality exists. Large groups of people sometimes make terrible collective mistakes because they reinforce one another's false beliefs and attach more value to their own solidarity around these beliefs than to such competing values as their own material welfare or that of their children. We cannot well understand that sort of danger if we cannot understand how beliefs and similar states could be part of the stuff of the world in the first place. *C&C* deserves to be acknowledged in the history of science as a landmark in the sequence of discoveries that enable that understanding. I also recommend it to those who want a brief nostalgic trip into Austinian Oxford philosophy on its deathbed, guided by one of its executioners.

References

Allori, V., Goldstein, S., Tumulka, R., & Zanghì, N. (2008). On the common structure of Bohmian mechanics and the Ghirardi–Rimini–Weber theory: Dedicated to Giancarlo Ghirardi on the occasion of his 70th birthday. *British Journal for the Philosophy of Science, 59*, 353–389.

Andersen, S., Harrison, G., Lau, M., & Rutström, E. (2014). Dual criteria decisions. *Journal of Economic Psychology, 41*, 101–113.

Atran, S. (1990). *Cognitive foundations of natural history*. Cambridge: Cambridge University Press.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New York: Wiley.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chomsky, N. (1959). A review of B. F. Skinner's verbal behavior. *Language, 35*, 26–58.

Churchland, P. (1979). *Scientific realism and the plasticity of mind*. Cambridge: Cambridge University Press.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67–90.

d'Espagnat, B. (2006). *On physics and philosophy*. Princeton: Princeton University Press.

Dennett, D. (1969). *Content and consciousness*. London: Routledge & Keegan Paul.
Dennett, D. (1978). Skinner skinned. In D. Dennett (Ed.), *Brainstorms* (pp. 53–70). Montgomery: Bradford.
Dennett, D. (1980). The milk of human intentionality. *Behavioral and Brain Sciences, 3*, 428–430.
Dennett, D. (1984). *Elbow room: Varieties of free will worth wanting*. Cambridge, MA: MIT Press.
Dennett, D. (1991a). *Consciousness explained*. Boston: Little Brown.
Dennett, D. (1991b). Real patterns. *Journal of Philosophy, 88*, 27–51.
Dennett, D. (1993). Back from the drawing board. In B. Dahlbom (Ed.), *Dennett and his critics* (pp. 203–235). Oxford: Blackwell.
Dennett, D. (2000). With a little help from my friends. In D. Ross, A. Brook, & D. Thompson (Eds.), *Dennett's philosophy: A comprehensive assessment* (pp. 327–388). Cambridge, MA: MIT Press.
Dennett, D. (2013). Kinds of things: Towards a bestiary of the manifest image. In D. Ross, J. Ladyman, & H. Kincaid (Eds.), *Scientific metaphysics* (pp. 96–107). Oxford: Oxford University Press.
Deutsch, D. (2010). Apart from universes. In S. Saunders, J. Barrett, A. Kent, & D. Wallace (Eds.), *Many worlds: Everett, quantum theory, and reality* (pp. 542–552). Oxford: Oxford University Press.
Deutsch, D. (2011). *The beginning of infinity*. London: Allen Lane.
Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive science. *Behavioral and Brain Sciences, 3*, 63–73.
Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive science: A critical analysis. In S. Pinker & J. Mahler (Eds.), *Connections and symbols* (pp. 3–71). Cambridge, MA: MIT Press.
Griffin, R., & Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.), *Daniel Dennett* (pp. 83–116). Cambridge: Cambridge University Press.
Hayek, F. (1952). *The sensory order*. Chicago: University of Chicago Press.
Hebb, D. (1949). *Organization of behavior*. New York: Wiley.
Humphreys, P. (2004). *Extending ourselves*. Oxford: Oxford University Press.
Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly, 32*, 127–136.
Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*, 507–531.
Kuhn, T. (1957). *The Copernican revolution*. Cambridge, MA: Harvard University Press.
Ladyman, J., & Ross, D. (2007). *Every thing must go: Metaphysics naturalized*. Oxford: Oxford University Press.
Ladyman, J., & Ross, D. (2013). The world in the data. In D. Ross, J. Ladyman, & H. Kincaid (Eds.), *Scientific metaphysics*. Oxford: Oxford University Press.
Melnyk, A. (2013). Can metaphysics be naturalized? And if so, how? In D. Ross, J. Ladyman, & H. Kincaid (Eds.), *Scientific metaphysics*. Oxford: Oxford University Press.
Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.
Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.
Reynolds, J., Hyland, B., & Wickens, J. (2001). A cellular mechanism of reward-related learning. *Nature, 413*, 67–70.
Rosenberg, A. (2011). *The Atheist's guide to reality*. New York: Norton.
Ross, D. (1994). Dennett's conceptual reform. *Behavior and Philosophy, 22*, 41–52.
Ross, D. (2000a). Rainforest realism: A Dennettian theory of existence. In D. Ross, A. Brook, & D. Thompson (Eds.), *Dennett's philosophy: A comprehensive assessment* (pp. 147–168). Cambridge, MA: MIT Press.
Ross, D. (2000b). The Dennettian stance. In D. Ross, A. Brook, & D. Thompson (Eds.), *Dennett's philosophy: A comprehensive assessment* (pp. 1–26). Cambridge, MA: MIT Press.
Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. Cambridge, MA: MIT Press.

Ross, D. (2011). Hayek's speculative psychology, the neuroscience of value estimation, and the basis of normative individualism. In L. Marsh (Ed.), *Hayek in mind: Hayek's philosophical psychology* (pp. 51–72). Bingley: Emerald.

Ross, D. (2013). Will scientific philosophy still be philosophy? In A. Zielinska (Ed.), *Repenser les rapports entre sciences et philosophie* (pp. 11–27). Grenoble: Recherches sur la philosophie et le langage.

Ross, D., Ladyman, J., & Kincaid, H. (Eds.). (2013). *Scientific metaphysics*. Oxford: Oxford University Press.

Seung, H. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron, 40*, 1063–1073.

Seyfarth, R., & Cheney, D. (2002). Dennett's contribution to research on the animal mind. In A. Brook & D. Ross (Eds.), *Daniel Dennett* (pp. 117–139). Cambridge: Cambridge University Press.

Sutton, R., & Bartow, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Vygotsky, L. (1934/1986). *Thought and language*. Cambridge, MA: MIT Press.

Zawidzki, T. (2007). *Dennett*. Oxford: Oneworld.

Zurek, W. (Ed.). (1990). *Complexity, entropy and the physics of information*. Boulder: Westview.

Chapter 3
What Was I Thinking? Dennett's *Content and Consciousness* and the Reality of Propositional Attitudes

Felipe De Brigard

> *Now once again is the view I am defending here a sort of instrumentalism or a sort of realism?*
> *I think that the view itself is clearer than either of the labels, so I will leave that question*
> *to anyone who still finds illumination in them.*
>
> (Dennett 1991)

Abstract Back in the 1980s and 1990s there was a lively debate in the philosophy of mind between realists and anti-realists about propositional attitudes. However, as I argue in this paper, both sides of this debate agreed on a basic assumption: that the truth (or falsehood) of our ascription of propositional attitudes has direct ontological implications four our theories about their nature. In the current paper I argue that such an assumption is false, and that Dennett had hinted at its falsehood in the first part of *Content and Consciousness*. In an exercise of "counterfactual exegesis", I suggest that, had this point been acknowledged then, this longstanding debate – which still survives to this date – could have probably been avoided.

Back in the 1980s and early 1990s, there was a lively debate in the philosophy of mind between realists and anti-realists about propositional attitudes. On the one hand, there was *intentional realism*, a view primarily defended by Jerry Fodor, who thought propositional attitudes were computational relations between a subject and a real, sentence-like representation in the language of thought. On the other hand, there were a handful of antirealist approaches, with Paul Churchland defending its most radical and influential version: *eliminative materialism*. For most empirically oriented philosophers of mind, this dispute is now obsolete, not so much because it has been settled, but rather because the field has evolved in such a way that many of the terms of the debate are no longer understood as they were back then. For

F. De Brigard (✉)
Center for Cognitive Neuroscience, Department of Philosophy,
Duke University, Durham, NC, USA
e-mail: felipe.debrigard@duke.edu

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_3

instance, mental representations are now rarely considered sentences in mentalese, and the few contemporary advocates of the language of thought support their views using cognitive and computational neuroscience, rather than using folk psychology as Fodor did (Gallistel and King 2009; Schneider 2011). Similarly, most views on computationalism have matured, and many no longer require the kinds of representational commitments Fodor once demanded (Piccinini 2008). However, in less empirically informed circles, this lack of denouement is taken to imply that the debate has simply remained dormant, and that the arguments deployed in the past are as strong now as they were before (see, for instance, Matthews 2010).

If only for that reason, my current attempt to revive a decades-old debate may not be completely futile. Yet, there is another reason why I think it is worth revisiting this dispute. I have long suspected that both intentional realists and eliminative materialists have based their arguments in a controversial thesis, viz. that the truth (or falsehood) of our ascriptions of propositional attitudes has direct ontological implications for our theories about their nature. This thesis, I believe, was underwritten by a particular take on scientific realism that committed both parties to accept two related assumptions: (1) that truth is as a matter of correspondence between words and things in the world, and (2) that the things named by true theories must exist. This sort of scientific realist stance was not ungrounded, of course. It was motivated by considerations regarding the success and failure of folk psychology. On the one hand, intentional realists took the *success* of our folk psychology as good evidence for the theory's truth, and then went on to suggest that our best theory of the mind should take the syntactic objects of our propositional attitudes as real entities – specifically, mental representations realized in the brain. On the other hand, eliminative materialists like Churchland took the relative *failure* of folk psychology as sufficient evidence for its falsehood, and then went on to suggest that folk psychology was false because it wrongly assumed the existence of unreal entities like beliefs, desires and so forth. The upshot of eliminative materialism was that, being a false theory, folk psychology was doomed to extinction, just like other obsolete theories we used to have.

As mentioned, this dichotomy largely framed the debate about the nature of propositional attitudes in the 1980s and 1990s (Fodor 1985). My contention now is that this was a false dichotomy, and that the debate was ill-construed. Moreover, I believe Daniel Dennett offered an important insight in the first part of *Content and Consciousness* (Dennett 1969; henceforth *C&C*) that, had it been developed, it would have severely weakened the aforementioned controversial thesis. Perhaps because Dennett did not develop this insight in the 1970s, and barely touched upon it when he further articulated his views on the nature of propositional attitudes (e.g., Dennett 1978, 1987, 1991), this important insight went unnoticed. As such, the current essay could be seen as an exercise in "counterfactual exegesis", as I try to develop this Dennettian insight in my own terms, writing on a line of argument that could have been explored years ago, and that might have prevented the development of a debate that, for many, it is now passé. Still, I hope that by incorporating some recent developments in related areas of philosophical research, those philosophers for whom the debate about the reality of propositional attitudes is merely dormant can find new reasons to question its legitimacy.

To that end, I offer an argument in which both eliminative materialists and intentional realists about propositional attitudes turn out to be partially wrong. Briefly stated, the idea is that these views represent two cardinally opposed ways of deriving ontological implications from the same underlying scientific realist assumption, which – I suggest – we would be better off rejecting. In order to make my case, I begin by explaining the origins of the dispute between intentional realists and eliminative materialists. I claim that it spawns from disagreements about a single argument – an argument I dub (inspired by Kitcher 2001) the *success-to-truth argument*. In Sect. 3.2, I talk about eliminative materialism. I argue that Churchland's arguments that folk psychology is false are unsound. I claim then that since there is no good reason to believe that folk psychology is false, the thesis of eliminative materialism cannot really get off the ground. In Part 3, I move on to a critical discussion about intentional realism. My criticism here is two-fold. On the one hand, on the basis of recent developments in linguistics and philosophy, I argue that we do not have enough a priori reasons to believe in the reality of 'that'-clauses' referents. On the other hand, I suggest that Fodor's inference to the best explanation vis-à-vis the reality of language-like mental representations can be challenged as well, casting more doubts on its ontological implications. Finally, in Sect. 3.4, I show how Dennett's insight in *C&C* can be read as anticipating these points, and as offering an alternative strategy to interpret *the success-to-truth argument*, in a way that might relieve the philosopher of mind from awkward ontological commitments regarding the nature of propositional attitudes.

3.1 The Success to Truth Argument

This is the formulation of what I call the success-to-truth argument (STA):

(Assumption) Folk psychology is a theory
(P1) Folk psychology is a successful theory
(P2) If a theory is successful, then it is true. Therefore,
(C1) Folk psychology is true.

Each statement needs some explaining. The Assumption holds that the so-called *'theory'-theory* is true. Barring some idiosyncratic differences in its formulation, the 'theory'-theory can be seen as the conjunction of two claims – the first of which, it appears, is contained by the second (Lycan 2004). The first claim is that mental terms are explanatory; they were inserted into our language to help us predict and explain other people's behaviors. The second claim is that these mental terms perform their explanatory and predictive role in virtue of being part of a theory, a *folk* theory, commonly known as *folk psychology*.

Folk psychology can be first approached by way of an analogy. Folk psychology is to scientific (organized, systematic) psychology as folk physics is to scientific (organized, systematic) physics. As we grow up and learn to navigate the world, we begin to develop an understanding of the structure of everyday objects, about the

way in which they behave, how they react with each other or under different conditions, and so forth. In general, folk physics works pretty well. Parental teachings instruct us to estimate with accuracy the trajectory of a baseball, and to then catch or flee accordingly. Less friendly classrooms have taught us to pick out tree branches apt to resist the stress produced by the gravitational force acting upon our well-fed 7-year-old bodies. Thanks to experience, we accrue piles of physical folklore that help us in the business of explaining and predicting the behavior of good old middle-sized objects. Mutatis mutandis, when it comes to folk psychology. Repeated encounters with energetically voiced instructions teach us when it may be wise to cut it out and do as our mother wishes. And our occasional interactions with persons whose behaviors we deemed questionable rightly suggest that they follow some beliefs we do not share. Just as we live in a world packed with middle-sized objects, we also live in a world populated with people. Folk psychology is the understanding we develop to make sense of people's complex behaviors.

It is customary to trace the historical origins of folk psychology back to Sellars' celebrated myth of Jones (Sellars 1956/1963). Details aside, Sellars' fable conveys the idea that mental terms are theoretical terms inserted in our folk psychology to refer to inner, unobservable episodes of others' mental lives – episodes which, are *alleged* to be causally responsible for their overt and observable behavior. Whereas our Rylean ancestors' theoretical repertoire was limited to mere observational/dispositional expressions, Sellars tells us that "Jones develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes" (Sellars 1956/1963: 186). These unobservable 'inner episodes' are to be taken as the referents of the theoretical mental terms Jones uses to explain the rich mental life unreachable by the behaviorist. To sum up: the Assumption says that folk psychology is a theory; that just like any other scientific theory, it works in part by introducing theoretical terms; that our mental terms are those theoretical terms; and that, hypothetically, mental terms refer to inner episodes.

The first premise (P1) insists that folk psychology is a successful theory. This premise, in fact, is the Rubicon dividing eliminative materialists and intentional realists. On the one hand, intentional realists suspect that, for the most part, folk psychology works fine. In general, predictions and explanations couched in mental terms seem to work, their generalizations seem to apply to novel cases, and their exceptions seem to be somewhat easily explained away, either by the theory itself, or by pointing at some violation of a ceteris paribus clause. On the other hand, eliminative materialists take folk psychology to be a complete failure, a stagnant science at most, with all sorts of predictive and explanatory shortcomings. Arguments in favor and against (P1) are, therefore, the main topic of the next section.

Finally, the second premise (P2) corresponds to what Kitcher (2001: 177) calls "the success to truth inference". The motivation behind (P2) is the belief that if scientific success is systematic, nothing miraculous must be going on; scientific accomplishments must not to be cashed out in terms of repeated coincidences but – at least intuitively – in terms of truth. Many scientific realists take (P2) as an argument in favor of scientific realism as, allegedly, it is the only view that does not make the success of science look like a sheer collection of systematic miracles. But

if this was the only option, one would seem to face an unfortunate dilemma: either one must embrace scientific realism, or one must accept the preposterous thesis that the success of science is pure luck (Votsis 2004). I hope to show, in Sect. 3.4, that some ideas in *C&C* can be read as offering an alternative view upon which to build a rejection of (P2) and a solution to the realism/anti-realism debate about propositional attitudes.

3.2 The Persistence of Folk Psychology

Eliminative materialism, according to Churchland, "is the thesis that our commonsense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience" (1981: 67). The force of this view, I contend, stems from the rejection of (P1). Notice, however, that Churchland needs (P2) to be stronger than the version I provided. He needs the implication in (P2) to be a bi-conditional. As it stands, it may very well be possible for folk psychology to be an unsuccessful theory and yet still be true. After all, there are instances in which certain theories, accepted as true by the relevant scientific community, have failed to produce successful predictions.[1] So Churchland needs (P2) to read:

(P2*) A theory is successful if and only if it is true

This way, if he can prove that folk psychology is actually an unsuccessful theory, its falsehood will be warranted – that is, C1 would be false. To that effect he cites "three major empirical failings of folk psychology" (Churchland and Churchland 1998: 8 [but see also Churchland 1981, 1988]):

(a) Folk psychology cannot explain a considerable variety of psychological phenomena, including mental illness, dreams, and concept acquisition by prelinguistic children, amongst many others.

[1] Here's a possible example of a theory that hasn't produced successful predictions, not because of the falsity of its premises, but because scientists don't know yet how to apply it in experimental or practical situations. Consider Schrödinger's equation. Although it is sufficiently clear which mathematical outcomes could be expected from calculations involving it, some empirical interpretations of such calculations are either unclear or impracticable. Cramer (1988), for instance, suggested an interpretation of the nature of wave equations, such as Schrödinger's, according to which a mixture of real and imaginary numbers is required. The problem is that these complex variables – as the mixed numbers are often called – are written as ± numbers, by virtue of which there are always two possible solutions. Alas, when used in equations involving the behavior of a system in time, the change in sign is supposed to be understood as "reversing" the direction of time, and that – as far as I understand – is still not quite easily interpretable in terms of empirical success. This impossibility, however, purports no harm to the acceptance of the equation as being true, and I suspect there may be similar examples in other areas of physics, perhaps even beyond quantum mechanics.

(b) Folk psychology has remained unaltered for the past 2,500 years, showing no signs of development and many of stagnation.
(c) Folk psychology does seem difficult to integrate with the other disciplines in its theoretical vicinity, like physics, chemistry, biology, and physiology.

The upshot, then, is that folk psychology is unsuccessful and should be deemed as false.

Despite the appeal of these alleged empirical reasons, I think they can be contested. Let us begin with (a). The *main* moral we were supposed to draw from Sellars' myth of Jones was that mental terms were introduced in our folk psychology in order to help us explain the observable complex behavior of other people. More specifically, mental terms were supposed to contribute to the systematization of laws, the purpose of which was to explain and predict the observable behavior of other persons. Now, Churchland considers that folk psychological explanations fail on two grounds: (1) because their theoretical terms depict a "radically inadequate account of our internal activities" (Churchland 1981: 570), and (2) because they prove ineffective when applied to a subset of psychological phenomena (e.g. mental illness, sleep, etc.). However, rejecting folk psychology on the grounds of (1) does not seem fair once we realize that "our internal activities" was not its proprietary domain of evidence and explanation in the first place. When it comes to scientific explanations, it is always important to keep the notion of success relative to the kind of object over which its predictions and explanations are supposed to operate. And it seems clear that in the case of folk psychology these objects are persons. Mental states were never introduced into our folk psychological language in order to stand in place of neural events. It is true that Jones *hypothesized* that theoretical mental terms – perhaps because they *seem to be* referential terms – were supposed to refer to inner linguistic episodes. However, this consideration, as well as any other further considerations regarding the *nature* of such episodes, is going to be either gratuitous or dependent upon subsidiary hypotheses (e.g. that our inner mental life mirrors our overt linguistic life; that mental states are to be correlated with brain states; that there are not non-linguistic inner episodes causally responsible for overt utterances, etc.). If you want to claim that inner episodes are brain events you may provide these subsidiary hypotheses. Nonetheless, for the purpose of the effectiveness of the myth, you need not. For all Jones knows, dualism could be true, the extended cognition hypothesis could be true, in fact, people could even be zombies, and yet folk psychology would still be vindicated. Why? Because the assumption of mental terms – that is, of theoretical terms – serves *primarily* the purpose of systematization: "it provides connections among observables in the form of laws containing theoretical terms" (Hempel 1958/1965: 186). Theoretical terms in our laws are, as it were, operational shortcuts posited in place of a bunch of observational data, which are further used to infer observational conclusions there-from. They do not serve primarily a referential purpose. Therefore, as long as they serve *their* purpose within the laws, whether they fail to refer to our internal neural activities doesn't really matter.

By the same token, to reject folk psychology on the grounds of (2) does not seem reasonable either. Suppose we agree that we have always used mental terms to make

sense of people's behaviors. Now, insofar as we have used mental terms in *this* way, psychological explanations and predictions are actually quite successful.[2] In general we are good at interpreting someone else's needs and hopes, what to expect from them given what we know, or even what we don't know. Indeed, the success of folk psychology in everyday life is so ubiquitous that it is "practically invisible" (Fodor 1985: 3). It is true that, at times, our explanations at the folk psychological level seem to fail. But there are failures and there are *failures*. Suppose I ask you to meet me tomorrow at school at 3:00 pm. Suppose further that you say, 'Yes, I'll be there'. From that piece of information I infer that you have formed the desire to meet me at school tomorrow and that you have formed the belief that I will be there at 3:00 pm. Then I put belief and desire together and I predict the following action: that you will go to school tomorrow at 3:00 pm for our meeting. The prediction fails, alas: you forgot the date. What went wrong? Here one has (at least) two options: one can either blame the entire predictive apparatus (i.e. folk psychology), or one can simply argue that your obliviousness constitutes a violation to a tacit ceteris paribus clause. Blaming the entire apparatus of folk psychology on the basis of just one failure seems a bit exaggerated. For one, I can provide an explanation of the failure in terms of the very same theory: if you hadn't *forgotten* the date, my prediction would have worked just fine. Secondly, it is true that similar extrapolations have proved successful in the past (last Wednesday – remember? – you did actually make it to our appointment). Finally, I can also be confident that the new prediction I make right after I talk to you – and you apologized, swore this time you'd be there on time, etc. – is actually going to work, ceteris paribus of course. Then again, maybe the problem is that you may not like ceteris paribus clauses at all. Fair enough. However, if that is so, your concerns can be generalized across the board, for they may actually affect most of our scientific theories (including neuroscience!), not only folk psychology (see, for instance, Lange 2002).

Surely Churchland does not have *those* cases of failure in mind when he claims that folk psychology cannot accommodate certain phenomena. He has in mind *big* failures, like the case of epilepsy. But was this really a failure of folk psychology? It seems to me that epilepsy is merely an exceptional disturbance whose behavioral characteristics are "less psychological" than the prototypical folk psychological phenomena. It is not that epilepsy was not easily explainable by reference to folk psychology's ceteris paribus clauses; it is rather that it was a very odd behavior, like hiccups or somnambulism, and it just did not seem to be the product of typical psychological states. Perhaps that was *precisely* the reason why people introduced demonic possessions to explain epilepsy: since it was not part of the domain of characteristic behaviors folk psychology usually explained, a different discipline was required to do the job. It is true that theology failed to explain the phenomena and that now neuroscience can explain epilepsy all right. However, it is not clear to me how this achievement of neuroscience is supposed to harm the success of a folk theory for which epilepsy was not clearly a proprietary explanandum. For not being able to explain epilepsy in terms of demonic possessions, it is not psychology that should not be blamed, but theology!

[2] Dennett articulated this point, before Churchland's paper, in pieces like *True Believers: The Intentional Strategy and Why It Works* (Reprinted in Dennett 1987).

Similar points can be made regarding other cases of *big* failures Churchland mentions. Take dreams for instance. Dreams do not elicit typical overt behaviors. People rarely behave when they are dreaming. And when they do, their behavior is rarely elicited by any inner episode they are aware of – or, at least, that they could causally respond to in virtue of their content. In that regard, dreams do not seem to be proprietary explananda of folk psychology. Therefore, insofar as they do not belong to the domain upon which folk psychological explanations were supposed to operate, it is unfounded to use them as counterexamples. A similar conclusion can be found in Horgan and Woodward (1985: 402) for whom "There is no good reason, a priori, to expect that a theory like [Folk Psychology], designed primarily to explain common human actions in terms of beliefs, desires, and the like, should also account for phenomena having to do with visual perception, sleep, or complicated muscular coordination" (Horgan and Woodward 1985: 402).

What about (b)? There is a longstanding line of argumentation against the stagnation objection trying to show that, in reality, folk psychology has actually progressed in the past 2,000 years. To that effect, philosophers and psychologists have shown that psychology, at the social and personal levels, makes constant use of belief/desire talk in the process of pushing forward their research programs: "for instance, temperament seems to be more useful in predicting behavior than other sorts of personality traits, according to social psychology; short-term memory holds about seven 'chunks' of information, whether these are numbers or names or grocery items, according to cognitive psychology; and so on" (Schroeder 2006: 69). I think this line of argument is basically right; I'd just add one more point: folk psychology not only proves necessary to the process of concocting research programs but, *more importantly*, to the process of carrying out those programs. It seems undeniable that true ascriptions of mental states are necessary when interpreting and producing neuroscientific data in situ, both inside and outside of the lab. Neuroscientists ought to believe that their subjects' introspective reports are veridical no less than they should trust the word of their co-workers. These intersubjective data would be useless unless we had the network of folk psychology up and running.

Still, there is another reason to be skeptical about the force of (b). 'Development' is a tricky word. In what sense does a theory develop? If developing counts as fostering research programs, then – as Horgan and Woodward (1985) argued – folk psychology has clearly developed. On the other hand, if development means something like "refinement" of a theory's axioms and principles, then I agree: folk psychology hasn't shown that much of it. But then again this sort of "immobility" need not be a sign of failure. It may be a sign of proper functioning instead. If a theory constantly proves unsuccessful and does not undergo revisions and changes, it is right to accuse it of being a bad theory. But if a theory works just fine when it has to, why would we want it to change at all? Consider basic arithmetic. Nobody would reject basic arithmetic on the grounds that it has not undergone any significant changes in the last 2,000 years. Basic arithmetic – the primary school arithmetic that most people operate with – hasn't changed because it works just fine for most everyday tasks. A similar point can be made about folk physics. People keep making the same rough generalizations and predictions about middle-sized mundane objects on

the feeble basis of previous successful experiences; yet, so far as quotidian life goes, folk physics works alright and hasn't shown signs of severe alterations. The same goes, mutatis mutandis, for folk psychology.[3]

Let me conclude with a comment about (c). To being with, it seems unclear what the objection amounts to. For the objection to be *really* an objection against the success of folk psychology the following claim should be true: that if a theory A is not integrable to a theory (or a set of theories) B, then A is unsuccessful. Call this claim *the integrability condition*. But what is meant by "integration"? In his 1981 paper, Churchland equates "integration" with the idea that some natural sciences tend toward a "theoretical synthesis" with the physical sciences in which the categories of the former are successfully reduced to those of the latter. But, he says, "F[olk] P[sychology] is no part of this growing synthesis. Its intentional categories stand magnificently alone, without visible prospect of *reduction* to that larger corpus" (Churchland 1981: 75). And it is fair to assume that by "reduction" he means what he meant 2 years before, in his 1979 book: that a theory A is successfully reduced to a theory B so long as two conditions are met: (1) that we can provide a set of rules (so-called "bridge laws") according to which the terms in A are mapped onto terms of a subset of sentences in B, and (2) that the expressions in B which the terms of A were mapped onto are axioms of A (Churchland 1979: 81ff). That way, A will be "contained" in B, i.e. B will explain as much as A explains and more. However, several arguments in the philosophy of science should have convinced us by now that (1) is not the case for most – if not for all – (special) sciences, and that since (2) presupposes the success of (1), (2) may prove impractical as well.[4] Therefore, given

[3] A different concern is to accuse folk physics of being unable to solve puzzles in the domain of scientific (organized, systematic) physics. This is also an unfair claim. Scientific physics deals with highly idealized objects and situations whereas folk physics has a more mundane domain and a very different purpose. I think it would be a mistake to reject folk physics on the basis that its generalizations don't coincide with the generalizations of scientific (organized, systematic) physics. The same, I think, goes for folk psychology. As Andy Clark so eloquently put it once: "Folk psychology may not be playing the same game as scientific psychology, despite its deliberately provocative and misleading label" (1989).

[4] I have in mind the arguments in Fodor's "Special sciences" (1974). For instance, the latter, very briefly, goes like this: a successful reduction of the psychological law like

(1) $S_1x \rightarrow S_2x$

is achieved as long as we can provide bridge laws of the form

(2a) S_1x *iff* P_1x and
(2b) S_2x *iff* P_2x,

guaranteeing the reduction of the psychological predicates S1 and S2 to neurophysiologic predicates P1 and P2 in a law of the form

(3) $P_1x \rightarrow P_2x$.

Alas, this sort of reduction is impracticable because bridge laws connecting type-psychological predicates with type-neurophysiologic predicates are, if not impossible, highly improbable ("an accident on a cosmic scale"). At most, all we can get are correlations between type-psychological predicates with heterogeneous disjunctions of type-neurophysiologic predicates like

the correct rendering of *the integrability condition* (if a theory A isn't *reducible* to another theory B, then A is unsuccessful), and given the arguments against the tenability of such reductions, the acceptance of *the integrability condition* required for the success of (c) would force us to reject any theory that proves irreducible as unsuccessful. Sadly, that would include basically all special sciences (not only psychology, but also economics, sociology, and so forth) and some lower-level sciences, like ecology, biology and perhaps neurology. To argue that none of these sciences is successful is preposterous. Irreducibility just cannot be the mark of scientific failure.

Some may object at this point that I am being unfair, as Churchland soon realized that his "classical account of intertheoretic reduction appeared to be importantly mistaken", and offered some "necessary reparations" (Churchland 1985/1992; Churchland and Hooker 1985). Fair enough. I'm willing to assume, for the argument's sake, that his new account actually circumvents the difficulties mentioned above. Still, there is another reason to be suspicious of the idea that reducibility speaks in favor of the success of a theory. If the success of a science is to be accounted for in terms of its explanatory and predictive achievements, then a successful reduction should have a negative effect on the explanatory power of the reduced science. In other words, a reduced science can't provide a better answer for a certain question than its reducing science. But this is hardly the case with folk psychology. Often times, the kind of explanations users of folk psychology require are not neurological. Sometimes we demand historical explanations, or accounts in terms of the environment in which the subject is embedded, or even contrastive answers, as when we wonder why a person decided to do X as opposed to Y. Reductive accounts may be able to provide us with full-fledged elaborations of the neural underpinnings of those behaviors, but it isn't obvious that an answer couched in neurological terms is going to be always, and for every possible purpose, explanatorily satisfactory. We frequently demand explanations in folk psychological terms, regardless of whether we have reductive accounts of the terms being used. I don't think it is clear at all that every why-question we may raise in folk psychological terms is suitable to be satisfactorily answered in neurological terms. Thus, issues about irreducibility seem to be orthogonal to preoccupations about the theory's success.

---

(4) Sx *iff* P_1x or P_2x or … or P_nx

in which case the right side of the bi-conditional won't correspond to a natural-kind of neurophysiology. Ultimately, the reduced law that uses type-neurophysiologic predicates would look like

(5) P_1x or P_2x or … or $P_nx \rightarrow P'_1x$ or P'_2x or … or P'_nx

where Pi and P'i are nomologically related. The problem, however, is that if the identity relation in the bridge laws (like 4) isn't between natural-kinds, then they aren't laws. But if they aren't laws then (5) isn't a law either. And when no laws, no reduction. QED.

3.3 There May Not Be Beliefs After All

If you have been convinced by the considerations in the previous section, then you might think that the eliminative materialist does not have sound reasons for claiming that folk psychology is unsuccessful. In addition, if you consider the STA a valid argument, then you probably think that folk psychology is true. None of the above, however, gives you intentional realism yet. To that end, we still need one further argument, which may be called *the truth-to-existence-via-reference argument:*

(PP1) Folk psychology is true.
(PP2) The statements of folk psychology report propositional attitudes.
(PP3) Propositional attitudes are two-place relations between subjects and the referents of 'that'-clauses.
(PP4) All things considered, the best candidates we have for referents of 'that'-clauses are mental representations in the language of thought. Therefore,
(CC2) There are mental representations in the language of thought.

Again, each premise needs some clarification. (PP1) is the conclusion of *the success-to-truth argument* (i.e. (PP1)=(C1)). (PP2) is a traditional tenet that can be traced back at least to Russell's (1918) lectures on logical atomism. According to this claim, mental states are to be characterized as ascribing to a subject *S* an intentional verb *Vs* (such as 'believes', 'fears', 'hopes', etc.) and a certain proposition *p*. Propositional attitude reports, thus, conform to the following general form: '*S Vs* that *p*', examples of which are "John hopes that it is raining", "Anne believes that having a small wedding is fine" and "Mario cree que el tiempo en Nueva York se siente distinto". Because propositional attitude reports conform to this general form, many believe that propositional attitudes are better understood as two-place relations between a subject and a proposition, which is the referent of the 'that'-clause. Indeed, it is customary to regiment propositional attitude statements in the following form:

[PA] $(\exists S)\ (\exists p)\ (R(S,p))$

where '*S*' refers to a subject, '*p*' refers to whatever the referent of the sentential complement clause may be (usually a proposition), and 'R' refers to the relevant intentional relation between them (Fodor 1978/1981; Schiffer 1992). Such is the rationale behind (PP3). In support of (PP3) Fodor gives three reasons[5] (Fodor 1978/1981: 178–179):

(a) "It is intuitively plausible. 'Believes' looks like a two-place relation, and it would be nice if our theory of belief permitted us to save appearances".[6]

[5] As mentioned, I'm confining my notion of intentional realism to Fodorian sentential realism. Because of that, the arguments in favor of (P3) and (P4) are his. Alternative accounts supporting (P3) and (P4) are not going tobe considered. It may be possible that my arguments apply to them as well, but they need not.

[6] Fodor uses "belief" as an illustration, but he's actually talking about all propositional attitudes. As such, his claims are to be read as extending to all propositional attitudes, not only to beliefs.

(b) "Existential Generalization applies to the syntactic objects of verbs of propositional attitudes; from 'John believes it's raining' we can infer 'John believes something' and 'there is something that John believes'."
(c) "The only known alternative to the view that verbs of propositional attitudes express relations is that they are (semantically) 'fused' with their objects, and that view would seem to be hopeless."

The force of all these reasons comes from linguistic and philosophical analysis of propositional attitude talk. The assumptions that support them will be discussed, when I present my arguments against (a), (b) and (c). Finally, (PP4) is basically an inference to the best explanation. The suggestion is that once you take into account all the data a theory of propositional attitudes is supposed to account for, the best candidate we end up with is a theory according to which "propositional attitudes are relations between organisms and formulae in an internal language; between organisms and internal sentences, as it were" (Fodor 1978/1981: 187). I think this inference to the best explanation can be blocked as well. Let us move on, then, to the challenges.

The first challenge goes against the claim, conveyed by (PP2) – and (a) – that mental states can (and need) be characterized as embedded within 'that'-clauses. It has been pointed out (e.g. Ben-Yami 1997) that some bona fide sentences reporting mental states cannot be rendered into the canonical form of propositional attitude reports ([PA] above). Consider the following sentences (examples 1 and 3, from Ben-Yami 1997: 85):

1. I want to sleep
2. Andrew knows how to multiply six digit numbers mentally
3. I trust Joan

A typical suggestion is to offer alternative paraphrases for these sentences, such as:

1*. I desire that I am asleep
2*. Andrew knows that to multiply six digit numbers mentally one needs to φ.
3*. I believe that Joan is trustworthy

But notice that these forced paraphrases introduce several problems. 1*, for instance, sounds odd. And this is not only a problem for English, as a quick look at the same proposition in French and Spanish, for instance, dissuades us from that option.[7] It may be argued that in order to get the correct, paraphrasing some extra linguistic maneuvering may be required, not at the surface level, but at the level of their deep structure (viz., 'that'-clause in 1 involves an implicit subject). Perhaps that could solve the problem for these cases, but if so one would like to know why we want to force our mental state reports to fit a certain kind of structure. I know of

[7] Contrast 1 with its Spanish translation "Quiero dormir" and its odd rendering into a canonical form: "Quiero que yo esté dormido". Ditto for French: "Je veux dormir" versus "Je veux que je sois endormi".

no argument to that effect (and neither does Ben-Yami 1997: 85). In the absence of such an argument it is hard not to conclude that the theory may be forcing the maneuver.

A related worry could be raised regarding 2*. I take it that all 2 tells us is that within Andrew's abilities we can count that of multiplying six digit numbers mentally. However, 2* seems to imply that if one were to ask Andrew how to multiply six digit numbers mentally he would be able to give us an answer in terms of φ. But 2* could be false while 2 be true. After all, Andrew may not know how it is that he manages to multiply six digit numbers in his mind. He knows that he can do it, but he may not know how or why he can do it.[8] And, finally, the same worry goes for 3*. All 3 tells us is that I trust Joan. It says nothing as to whether I believe that Joan is trustworthy. I could still stubbornly trust Joan despite the fact that I am seriously suspicious about her trustworthiness. Finally, I think that these considerations also speak against the first reason Fodor offers in support of (PP3). If not all mental states' attributions are suitable to be translated into statements of the canonical [PA] form, those that are can only constitute a subset of folk psychological statements. So it is not true that all folk psychological statements are better seen as two-place relations, as Fodor suggests.[9]

For the sake of the argument, however, let's assume that it is, in fact, intuitively plausible to render all our attribution of mental states in the canonical [PA] form. That is, suppose we accept that mental states can be paraphrased without semantic loss as expressing a two-place relation between subjects and the referent of 'that'-clauses – whether as propositions in abstracta or, as in the case of Fodor, presumably as neural concreta. Does that constitute enough reason to believe that the referents of 'that'-clauses are real? The answer is *no*. More assumptions need to get accepted for that conclusion to follow. Fodor gives us two reasons in support of (b): first, that 'that'-clauses behave referentially, and second, that existential generalization applies to 'that'-clauses. Now: why are these two reasons a good argument in support of there being referents of 'that'-clauses? It seems to me (and I'm not alone; see Balaguer 1998) that what underwrites this claim is basically Quine's criterion of ontological commitment plus an "intentional" reading of the Quine-Putnam indispensability thesis. Let me elaborate by comparing the case at hand with that of mathematics. Due to the influence of the Quine-Putnam indispensability thesis[10] in

[8] Notice that this is *not* a problem of expressibility. It isn't that Andrew does not know how to put into words what he does; it is rather that he may have no idea how he does it – he may not even know how to *begin* explaining what he does.

[9] A recent movement in epistemology, often called *intellectualism*, argues that know-how is a species of know-that (e.g., Stanley and Williamson 2001). If this was the case, then, it would follow that know-that statements should be translatable without semantic loss into know-how statements. Although arguing against intellectualism goes beyond the scope of the current essay, it may be worth pointing out that it remains a very controversial proposal, one that a growing number of philosophers reject (e.g., Noë 2005; see Fantl 2008, for a review).

[10] The claim, roughly, that if one's best scientific (physical) theory [after regimentation onto first-order logic] requires existential quantification over certain entities, then one is ontologically committed to such entities (Azzouni 1998: 1).

mathematics, theoretical irreducibility (and non-eliminability) is often assumed to carry with it ontological commitment. For it is frequently accepted that if *S* is irreducible to *R* (=$_{df}$ untranslatable to the other via bridge laws [see footnote 4]) and, when regimented, both *Sr* and *Rr* turn out to quantify over different variables,[11] then one is *eo ipso* committed to the existence of those entities (or kind of entities) picked up by the bound variables. In the case of mathematics such is the case with numbers (sets). I contend that for (b) to count as ontologically significant, the same should go for propositional attitudes (see also Balaguer 1998).

This argumentative line could be blocked with two moves. The first move is to show that 'that'-clauses do not behave referentially. The second move is to show that although existential generalization applies to 'that'-clauses, such a quantificational device can be read as being ontologically innocent, i.e. as conveying no ontological commitments by itself.

Let us begin with the first move. In general, objections against the non-referentiality of 'that'-clauses have been directed toward theories holding that the referents of 'that'-clauses are propositions. I believe that the force of at least two of these objections carry over to Fodor's analysis of propositional attitudes as being relational. The first of these objections in known as *the substitution failure*. Briefly stated the substitution failure objection says that if 'that'-clauses were really referential, and if their referents were really propositions, then they should share their denotations with linguistic constructions of the sort "the proposition that *p*" (Moltmann 2003: 82ff). However, this sort of substitution often fails. Consider the following substitution case:

4. John fears that Palin will be our next president.
5. John fears the proposition that Palin will be our next president.

Ex hipothesi, "that Palin will be our next president" and "the proposition that Palin will be our next president" share their reference: namely, the proposition that says that Palin will be our next president. But to be afraid of the eventual situation of Palin being the next president is different from being afraid of a proposition. It seems obvious that 4 and 5 differ in truth-value, so we should better conclude that 'that'-clauses do not refer to propositions (Hofweber 2006b). Now, does this concern carry over when we aren't talking about abstracta but concrete sentences in the language of thought? Consider:

6. John fears the mental sentence that Palin will be our next president.

Would 6 change the outcome of the substitution failure objection? I'm afraid not, at least insofar as the substitution failure objection counts as an argument *against* the relational analysis of propositional attitude reports. In order for (b) to count as a

[11] "Turn out" is short for: Take *Px* to be a formula with a free variable *x*, and take ∃ *(x)(Px)* to be directly deducible from S_r but not from R_r. Given Quine's criterion for ontological commitment, one is here committed to the existence of the referent of the variable in *Px* bound by the existential quantifier. Now: take ∃*(x)(Qx)* to be deducible from R_r but not from S_r. I take that if the criterion is correct, then it "turns out" that one is committed also to the existence of the referent of the variable in *Qx* bound by the quantifier (All under the assumption that one can have regimented versions of both *S* and *R*, my S_r and R_r Quine 1948).

linguistically valid reason in favor of 'that'-clauses being referential, Fodor needs that whatever goes for propositions goes too for mental formulae. And he cannot argue in favor of the latter as opposed to the former on the basis of some property that mental formulae but not propositions may possess. Remember that Fodor wants 'that'-clauses to be referential so he can claim, a priori, that there *must be* referents of 'that'-clauses. Using an alleged property about their nature to justify the argument in favor of their existence is circular.

The second objection I have in mind against 'that'-clauses being referential is originally due to Kripke (1979), although more recently has been developed by Bach (1997). The relational analysis of propositional attitudes finds support partly because it seems to reflect the apparent logical form of inferences like:

I1: A believes that *p*
B believes that *p*
→ There is something that both A and B believe.

However, when Kripke introduced his Paderewski-case puzzle he showed us that inferences of the form I1 aren't always valid. Suppose Carl meets Paderewski at a business meeting and as a result fixes the belief that Paderewski is a nice guy. Carl is pretty bad with faces, though. Later on he comes across Paderewski at a cocktail party where Paderewski strikes him as an annoying guy. As a result he forms the belief that Paderewski is not a nice guy. If the relational account of propositional attitude reports is correct, it seems as though Carl believes contradictory things. Specifically,

I2: Carl believes that Paderewski is a nice guy.
Carl disbelieves that Paderewski is a nice guy.
→ There is something that Carl both believes and disbelieves.

But Carl isn't being irrational; he's just ignorant about the fact that he's taking the name "Paderewski" to refer to two distinct individuals. Notice, however, that this fact is inessential to the problem. As Bach notes, when it comes to the relational analysis of propositional attitude reports, the believer need not have "any familiarity with the name in question or have any name at all for the object of belief" (Bach 1997: 224). Consequently, it seems that the two premises in I2 have Carl believing and disbelieving different things. If so, then I2 is not a valid inference. But given the fact that there aren't relevant formal differences between I1 and I2, we have no reason to believe that the linguistic appearances in I1 aren't misleading as well. To solve the puzzle Bach suggests that we reject an essential ingredient of the relational analysis of propositional attitude ascriptions: the assumption "that the 'that'-clause in a belief report specifies the thing that the believer must believe if the belief report is to be true" (Bach 1997: 221). In his account, 'that'-clauses *describe* their content instead (i.e., purport to state their content under a certain description, which may or may not be incomplete). Without this assumption, we have very little reason to take 'that'-clauses as referential.

Fodor can reject Bach's solution and stick to a relational analysis under the assumption that 'that'-clauses refer to mental sentences, which, unlike propositions, are neither ambiguous nor semantically incomplete. But this would be an unjustified

move. Remember that (b) – and for that matter (PP3) – was supposed to convey pre-theoretical reasons in favor of 'that'-clauses being referential. Latching onto alleged properties of hypothesized mental sentences to save the linguistic phenomena whose clarity was supposed to motivate the relational analysis in the first place is question begging.[12]

Still, there is a further motivation to reject (b). Even if one accepts that 'that'-clauses are referential, the only reason Fodor seems to offer to jump from that linguistic fact to the conclusion that their referents exist is a commitment to an ontologically loaded reading of existential generalization. Since belief reports admit of existential generalization ranging over their 'that'-clauses (e.g., the example in I1), and since 'that'-clauses admit no reduction to another language whose ontological commitments we could be more comfortable with ("Behaviorists used to think such translations might be forthcoming, but they were wrong" [Fodor 1978], see also footnote 6), then we *should* go ahead, as Quine taught us, and accept the referents of 'that'-clauses as real (Quine 1948; see also Fodor 1987: 15).

Why would Fodor want us to do this? He cannot be suggesting this move on the basis of his acceptance of Quine's theory of reference; after all, Fodor is known for his rejection of Quine's holism tout court. A more plausible answer is that he is doing so on the basis of a weaker assumption: that the best – if not the only – way to understand existential generalization is by treating it as ranging over domain-independent entities. But this is a contentious claim. One can instead adopt what Hofweber calls "an internalist view" about quantification and deem existential generalization as a logical device to increase expressive power, and a logical tool that allows us to talk about infinitary disjunctions of single instances (Hofweber 2006a) – which is in this case, infinitary disjunctions of instances of attributions of mental states. If so, then, existential generalizations would be ontologically innocent.[13] The internalist view of existential generalization could turn out to be wrong, of course, but it is a good alternative. And without an argument against it – or without an argument in favor of a domain-independent reading of quantification – we would be better off remaining agnostic as to whether we should take existential generalizations as unquestioned carriers of the ontological burden of our regimented theories. As Jody Azzouni pointed out – in a rather different context – without an independent argument of that sort, it seems that the only reason we have to take the ordinary phrase "there is/are" to commit us to the existence of whatever it seems to commit us to, is simply "that the ordinary language 'there is' *already* carries ontological weight" (Azzouni 1998: 4). Does Fodor have an argument in favor of the reality of propositional attitudes independent of an ontologically loaded reading of existential quantification? He sure does – that's the bulk of the argument for (PP4).

Before we switch toward that discussion, however, let me say something very briefly about reason (c) for (PP3). In light of the previous considerations, it may be

[12] If we allow the resources of a theory to explain this phenomenon, a connectionist approach sensitive to graceful degradation and assignment by omission may turn out to do a better job than the language of thought when it comes to explaining why Carl forgot Paderewski's face to begin with.

[13] Free logic also allows to read existential quantifiers as ontologically innocent (Orenstein 1990).

clear that the force of (c) has now diminished. Fodor's original rejection of the "fusion" theory was supposed to mobilize the intuition that *unlike* that theory, a relational account of propositional attitudes faced no problems. But we have seen that relational accounts face severe objections too. Indeed, contemporary attempts to explain away precisely those objections seem to favor instead non-relational accounts of propositional attitude reports (see, e.g., Moltmann 2003, for a neo-Russellian account, as well as the appendix of that paper for other non-relational alternatives). Consequently, even if the fusion theory is false, we still need more reason to prefer a problematic relational account.

So what about (PP4)? Truth be told, Fodor can accept all the aforementioned objections and reject (PP3), and still argue in favor of his intentional realism on the grounds of (PP4) alone. He may say that, *all things considered*, intentional realism constitutes the best *empirical* theory we have to "vindicate" – his word – folk psychology. That is, he may well accept that we do not have either linguistic or a priori metaphysical reasons to accept the reality of sentence-like mental states, and still hold that such a hypothesis needs to be accepted on empirical grounds. At the end of the day, this has been his preferred strategy. Sheltered by the motto "the only game in town", the hypothesis of the language of thought has been advertised as the best theory we can muster to explain several psychological phenomena. Niceties aside, his argument boils down to an inference to the best explanation for some puzzling phenomena: concept acquisition, the compositional, systematic, and productive character of our thought, the projectability of mental terms in our psychological laws, and some (but not very many!) more. Copious pages have been written in an attempt to provide alternative accounts of these phenomena in terms that do not force us to accept a language of thought (see, for instance, Jackendoff 1992; Millikan 1984; Prinz 2002; Fodor 1990). I'm afraid I will not contribute to the discussion. Instead, I am going to try a different tack.

If Fodor's argument for the truth of intentional realism boils down to an inference to the best explanation, then it had better be the case that an inference to the best explanation constitutes a *good* reasoning pattern for realism about theoretical or unobservable entities. After all, folk psychology is just another theory – unrefined if you want, and operational over a slightly different domain than scientific psychology – but a theory none-the-less. Recall that folk psychology's mental terms are theoretical expressions whose alleged referents are unobservable inner episodes, i.e. mental states. Now, scientific realists usually take inferences to the best explanation as good argumentative patterns in favor of the truth of a certain theoretical hypothesis. In brief, the rationale behind the inference to the best explanation is that if a certain hypothesis *H* explains a certain phenomenon X better than any of its rival hypothesis, then *H*'s explanatory superiority should be taken as a mark of its truth – or, at least, as a mark of its approximate truth. From there, however, scientific realists often jump to the conclusion that the unobservable entities postulated by the theory must be real. Fodor, as we have seen, is no exception here. He takes the hypothesis of the language of thought to be the best hypothesis we have to account for the aforementioned psychological phenomena, and then goes on to claim that this is enough reason to believe that it is true that there are sentence-like representations in our brains.

Notwithstanding the widespread use of inferences to the best explanation by scientific realists, its validity as an argument to support the truth of a scientific hypothesis has been challenged on several grounds. Perhaps the most common attack comes from scientific anti-realism. To begin with, scientific anti-realists – like Bas van Fraassen (1980) and Nancy Cartwright (1983) – have argued that being a good hypothesis is never enough ground for believing that it is true. After all, the set of all rival hypotheses we can choose from may contain only false ones. Moreover, as van Fraassen remarked (1980: 21ff), when a scientist is in the business of accounting for some observational evidence, she does not really choose the best possible explanation *there is*, but rather the best explanation that is available to her. It would be a mistake to infer from that fact that such a hypothesis must be true, or closer to the truth than any other hypothesis she may or may not have access to.

Furthermore, van Fraassen also noted that most scientific realists take the thesis of scientific realism *itself* as an inference to the best explanation, insofar as it is the best hypothesis we can muster to explain the success of science (see Fine 1984). According to them, the success of a theory mustn't be cashed out in terms of sheer luck. Scientific realism is the best hypothesis we have to reject that preposterous conclusion. Now, the circularity of the maneuver isn't worrisome, yet it opens the door for a rival hypothesis to scientific realism, namely that "we are always willing to believe that the theory that best explains the evidence, is empirically adequate (that all the observable phenomena are as the theory says they are)" (van Fraassen 1980: 20). This anti-realist alternative to scientific realism, known as *constructive empiricism*, tells us that if a theory is successful, then it is empirically adequate, and that a theory is empirically adequate "exactly if what it says about the observable things and events in this world, is true – exactly if it 'saves the phenomena'" (van Fraassen 1980: 12).

My tactic to reject (PP4) should be obvious now; if Fodor's argument for intentional realism boils down to no more than an inference to the best explanation, and if inferences to the best explanation aren't conclusive reasons to believe in the reality of postulated entities, then (PP4) does not constitute a conclusive reason to infer the existence of mental formulae coded in our brains. With the previous arguments against (PP2) and (PP3), I tried to show that the jump from truth to existence *via* reference depended solely on the viability of inferences to the best explanations as valid arguments for the existence of unobservable entities. But as we just saw, inferences to the best explanation do not provide such conclusive grounds. Even if *all things considered* the language of thought turns out to be the best hypothesis we have to explain some behavioral (i.e. observational) phenomena, it is still unwarranted to infer that there *are* mental formulae in our brain. Again, I'm *not* saying that the hypothesis of the language of thought is false. All I'm saying is that the *truth-to-existence-via-reference argument* won't get us from the truth of our ascriptions of propositional attitudes to the reality of mental formulae in our brains. Which is why, I think, the best strategy for the metaphysically cautious philosopher of mind seeking to understand the place of propositional attitudes in our ontological repertoire is to approach the issue from an ontologically innocent anti-realist perspective (perhaps akin to constructive empiricism), and to proceed gradually,

studying each propositional attitude ascription in its context of occurrence, the events – both behavioral and neural – with which they correlate, while taking as real only those parts of the explanations we have empirical evidence for.

3.4 Dennett's 'Prefutation' in C&C

To recap: In Sect. 3.1, I introduced the *success-to-truth argument* and suggested that both eliminative materialism and intentional realism spawned from different takes on it. In Sect. 3.2, I argued against Churchland's reasons to consider folk psychology unsuccessful. Finally, in Sect. 3.3, I presented some objections against the *truth-to-existence-via-reference argument* in order to prove it insufficient to support intentional realism. In the end I defended a metaphysically innocent approach toward propositional attitudes, very much in the spirit of van Fraassen's constructive empiricism, according to which our ontological commitments to the mental entities mentioned in our propositional attitude ascriptions should proceed in conformity with our empirical evidence in favor of their existence.

This is precisely Dennett's insight in *C&C*. He came to it from a different perspective, of course; he was arguing for the non-referentiality of mental terms and the plausibility of a fusion-view, according to which intentional statements should be taken as wholes when it comes to evaluating their truth values. However, his endorsement of the fusion-view was, at best, half-hearted. His real motivation, I believe, was to convince us that in order to advance the discussion about the reality of mental terms, we needed to temporarily withhold our grammatically driven metaphysical assumptions, at least until we reached a clearer understanding of the nature of the phenomenon whose reality is supposed to be at stake. His *tentative fusion* approach is, in this sense, methodological:

> We wish to proceed with no ontological presuppositions to the effect that mental entity terms either are or are not referential, and this can be accomplished by treating all sentences containing mental entity terms as tentatively fused, subject to further discoveries which will lead us to confirm the fusion or relax it. (*C&C*, 16)

Notice that the metaphysical innocence with which Dennett thinks intentional statements should be approached does not prevent him from regarding them as truth-evaluable:

> In most general terms our task is to provide a scientific explanation of the differences and similarities in what is the case in virtue of which different mental language sentences are true and false. Thus, for example, our task is not to identify Tom's thought of Spain with some physical state of his brain, but to pinpoint those conditions that can be relied upon to render the whole sentence 'Tom is thinking of Spain' true or false. This way of proceeding still characterizes the task of finding an explanation of the mind which is unified with, consistent with, indeed a part of science as a whole, but eschews—at least initially—the obligation to find among the things of science any referents for the terms in the mental vocabulary. (*C&C*, 18)

At this juncture, I think it is useful to see Dennett's view as a sort of re-interpretation of Sellars' myth of Jones in the spirit of van Fraassen's constructive empiricism.

Recall that, according to Sellars, back in the days of the mythical Jones, our Rylean ancestors were Positivists as well. They believed in a difference between observational and theoretical terms, according to which the former referred to observable entities and the latter to unobservable entities. But this dichotomy, as van Fraassen (1981) showed us, conflates two different distinctions: the distinction between observational and theoretical *terms*, on the one hand, and observable and unobservable *entities*, on the other. Whether or not an entity is observable has nothing to do with language: it has to do with observation. Accordingly, it is a mistake to think that because intentional terms got into our folk psychological language as theoretical terms, they must refer to entities that are unobservable, either in principle (e.g., states of the soul), or in practice (e.g., states of the brain).

Similarly, Dennett points out that the fact that our intentional terms appear to behave referentially does not necessarily mean that they must refer to some kind of unobservable entity, stuck in the middle of a causal chain of observable entities, and ontologically on par with them. Thus, he writes:

> So, one can only ascribe content to a neural event, state or structure when it is a link in a demonstrably appropriate chain between the afferent and the efferent. The content one ascribes to an event, state or structure is not, then, an extra feature that one discovers in it, a feature which, along with its other, extensionally characterized features, allows one to make predictions. Rather, the relation between Intentional descriptions of events, states or structures (as signals that carry certain messages or memory traces with certain contents) and extensional descriptions of them is one of further interpretation. […] The ideal picture, then, is of content being ascribed to structures, events and states in the brain on the basis of a determination of origins in stimulation and eventual appropriate behavioral effects, such ascriptions being essentially a heuristic overlay on the extensional theory rather than intervening variables of the theory. (*C&C*, 78–80)

Needless to say, the idea that we ascribe intentional states to others – as when we attribute propositional attitudes to them – as a heuristic to make sense of their behaviors (both afferent and efferent) became the pillar of what is oftentimes called the "instrumentalism" of the *intentional stance* (Dennett 1978, 1987). What I find surprising, having read *C&C* after studying much of what went on with the intentional stance in the 1980s and 1990s, is that critics typically accused Dennett of not respecting the ontological commitments that truth-bearing ascriptions of intentional statements, such as propositional attitudes, carry with them. To put it simply: critics thought that if he wanted propositional attitudes ascriptions to be truth-evaluable, then he had to take a stand regarding their reality. More precisely, critics thought that he either had to be committed to some sort of intentional realism if propositional attitude reports were to come out true, or some sort of eliminativism if they were to come out false. But Dennett didn't have to. He argued in *C&C* that whether a particular propositional attitude ascription comes out as true is independent of whether the intentional term embedded in it picks out something concrete in the brain (or in the soul). And this, I contend, amounts to a prefutation – i.e. a Dennettism meaning a refutation that is offered before an argument is raised (Dennett 1996) – of the claim that the truth (or falsehood) of our ascriptions of propositional attitudes carry ontological weight onto our theories about the nature of mental states – a widely shared but mistaken assumption in the realism/antirealism debate of the 1980s and 1990s about propositional attitudes.

I hope that this essay helps to place some arguments found in *C&C* within the context of the contemporary debate about truth ascription and ontology as it relates to intentional statements. No doubt there is much more that could be said about the relationship between truth ascriptions to intentional statements and the reality of propositional attitudes within Dennett's system. For instance, I think it might be worth exploring the extent to which the intentional stance can latch onto the theoretical resources offered by constructive empiricism when it comes to issues such as the reality of propositional attitudes. On the face of it, its seems like a relatively straightforward task. Traditionally, constructive empiricism and deflationism about truth have been lumped together. Given Dennett's Quinean inclinations it wouldn't be surprising if a constructive empiricist reading of his instrumentalism would end up supporting a deflationist view on the truth of propositional attitude ascriptions. However, recent developments suggest otherwise. As Jamin Asay (2009, 2012) has recently argued, constructive empiricism requires a more substantive theory of truth than deflationism. Would the same be the case for Dennett's view? In other words, does Dennett's instrumentalism require a more substantive view of truth than deflationism? If so, would that conflict with other Quinean aspects of his philosophy? And, what would be then the best truthmaking theory for Dennett's instrumentalism? These, I believe, are all questions worth asking, although their answers might have to wait for another day, and maybe for someone else.[14]

References

Asay, J. (2009). Constructive empiricism and deflationary truth. *Philosophy of Science, 76*(4), 423–443.

Asay, J. (2012). A truthmaking account of realism and anti-realism. *Pacific Philosophical Quarterly, 93*(3), 373–394.

Azzouni, J. (1998). On 'On what there is'. *Pacific Philosophical Quarterly, 79*, 1–18.

Bach, K. (1997). Do belief reports report beliefs? *Pacific Philosophical Quarterly, 78*, 215–241.

[14] A final, personal note: I read *C&C* for the first time in the summer of 2006. It was part of my background reading toward writing my MA thesis on the nature of propositional attitudes. I had read Dennett's work before, but never *C&C*. It also happened that, as soon as I finished part I of *C&C*, I went on to sail with Dennett and others on his boat *Xanthippe*, and at some point the subject of *C&C* emerged. 'Have you read it?' Dan asked. I told him that I had just finished the first part. 'And what did you think?' You see, at the time, I was not only working on my thesis; I was also working on my English, and my answer did not come across as intended. 'I was disappointed', I said, and laughter ensued. But what I meant to say is that I was disappointed to see that what I thought was an original idea in my MA thesis, turned out to have been there, masterfully articulated, in the first chapters of *C&C*! I did not abandon the project though, for notwithstanding the parallelisms between the claims in *C&C* and mine, I still thought it was worth showing how one could arrive at the same conclusion through a different path – this, I guess, is philosophy's way of reaching convergent evidence. Thus, the present essay draws heavily from my MA thesis, and I hope it helps to clarify my poor choice of words back when we were on *Xanthippe*! I also would like to thank the following people for their helpful comments on previous drafts: Jamin Asay, Jody Azzouni, Max Beninger, Alex DeForge, Dan Dennett, Anne Harris, Thomas Hofweber, Joshua Knobe, Gualtiero Piccinini, Jesse Prinz, and Kate Ritchie.

Balaguer, M. (1998). Attitudes without propositions. *Philosophy and Phenomenological Research, 58*(4), 805–826.
Ben-Yami, H. (1997). Against characterizing mental states as propositional attitudes. *The Philosophical Quarterly, 47*(186), 84–89.
Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon.
Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge: Cambridge University Press.
Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, *78*(2). Reprinted in: Churchland, 1992, 1–22.
Churchland, P. M. (1985). Reduction, qualia, and direct introspection of brain states. *Journal of Philosophy*, *82*(1). Reprinted in: Churchland, 1992, 47–85.
Churchland, P. M. (1988). *Matter and consciousness*. Cambridge, MA: MIT Press.
Churchland, P. M. (1992). *A neurocomputational perspective*. Cambridge, MA: MIT Press.
Churchland, P. M., & Churchland, P. S. (1998). *On the contrary: Critical essays*. Cambridge: MIT Press.
Churchland, P. M., & Hooker, C. A. (1985). *Images of science: Essays on realism and empiricism*. Chicago: University of Chicago Press.
Clark, A. (1989). *Microcognition*. Cambridge, MA: MIT Press.
Cramer, J. G. (1988). An overview of the transactional interpretation. *International Journal of Theoretical Physics, 27*, 227.
Dennett, D. C. (1969). *Content and consciousness*. New York: Routledge.
Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dennett, D. C. (1991). Real patterns. *Journal of Philosophy, 88*, 27–51.
Dennett, D. C. (1996). Did HAL commit murder? In D. G. Stork (Ed.), *Hal's legacy: 2001's computer as dream and reality*. Cambridge, MA: MIT Press.
Fantl, J. (2008). Knowing-how and knowing-that. *Philosophy Compass, 3*(3), 451–470.
Fine, A. (1984). The natural ontological attitude. In J. Leplin (Ed.), *Scientific realism*. Berkeley: University of California Press.
Fodor, J. (1974). Special sciences and the disunity of science as a working hypothesis. *Synthese, 28*, 97–115.
Fodor, J. (1978). Propositional attitudes. *The Monist, 64*(4), 501–524.
Fodor, J. (1981). *Representations*. Cambridge, MA: MIT Press.
Fodor, J. (1985). Fodor's guide to mental representation. *Mind*, Spring, 66–97.
Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
Fodor, J. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.
Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain*. Chichester/Malden: Wiley.
Hempel, C. G. (1958). *The theoretician's dilemma: A study in the logic of theory construction* [Reprinted in: *Aspects of scientific explanation and other essays in the philosophy of science* (1965)]. New York: Free Press.
Hofweber, T. (2006a). Schiffer's new theory of propositions. *Philosophy and Phenomenological Research, 73*, 211–217.
Hofweber, T. (2006b). Inexpressible properties and propositions. In D. Zimmerman (Ed.), *Oxford studies in metaphysics* (Vol. 2). Oxford: Oxford University Press.
Horgan, T., & Woodward, J. (1985). Folk psychology is here to stay. *The Philosophical Review, 44*(2), 197–226.
Jackendoff, R. (1992). *Languages of the mind*. Cambridge, MA: MIT Press.
Kitcher, P. (2001). Real realism: The galilean strategy. *Philosophical Review, 110*(2), 151–197.
Kripke, S. (1979). A puzzle about belief. In A. Margalit (Ed.), *Meaning and use*. Dordrecht: D. Riedel.
Lange, M. (2002). Who's afraid of Ceteris-Paribus laws? Or: How I learned to stop worrying and love them. *Erkenntnis, 57*(3), 407–423.

Lycan, W. (2004). *Eliminativism*. (Unpublished) Available at: http://www.unc.edu/%7Eujanel/3255H5.htm
Matthews, R. (2010). *The measure of mind*. Oxford: Oxford University Press.
Millikan, R. G. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.
Moltmann, F. (2003). Propositional attitudes without propositions. *Synthese, 135*(1), 77–118.
Noë, A. (2005). Against intellectualism. *Analysis, 65*(4), 278–290.
Orenstein, A. (1990). Is existence what existential quantification expresses? In R. B. Barrett & R. F. Gibson (Eds.), *Perspectives on quine* (pp. 245–270). Cambridge: Blackwell.
Piccinini, G. (2008). Computation without representation. *Philosophical Studies, 137*(2), 205–241.
Prinz, J. (2002). *Furnishing the Mind*. Cambridge, MA: MIT Press.
Quine, W. V. O. (1948). On what there is. *Review of Metaphysics, 2*, 21–38.
Russell, B. (1918). *The philosophy of logical atomism* [Reprinted in Pears, D. (1985)]. Chicago: Open Court.
Schiffer, S. (1992). Belief ascription. *The Journal of Philosophy, 89*, 499–521.
Schneider, S. (2011). *The language of thought: A new philosophical direction*. Cambridge, MA: MIT Press.
Schroeder, T. (2006). Propositional attitudes. *Philosophy Compass, 1*(1), 56–73.
Sellars, W. (1956). *Empiricism and the philosophy of mind* [Reprinted in: *Science, perception and reality* (1963)]. New York: Routledge & Kegan Paul Ltd.
Stanley, J., & Williamson, T. (2001). Knowing how. *Journal of Philosophy, 98*(8), 411–444.
Van Fraassen, B. (1980). *The scientific image*. Oxford: Oxford University Press.
van Fraassen, B. (1981). Critical study: Paul Churchland, scientific realism and the plasticity of mind. *Canadian Journal of Philosophy, 11*, 555–567.
Votsis, I. (2004). *The epistemological status of scientific theories: An investigation of the structural realist account*. Ph.D. dissertation, London School of Economics.

Chapter 4
Dennett's Dual-Process Theory of Reasoning

Keith Frankish

Abstract *Content and Consciousness* (*C&C*) outlines a framework for thinking about the relation between mind and brain that has been hugely influential and salutary. This chapter discusses a relatively neglected aspect of this framework – the treatment of thinking and reasoning in Chap. VIII. Here Dennett distinguishes two senses of "thinking", parallel to the senses of "awareness" distinguished earlier in the book. In one sense "thinking" refers to sub-personal information processing whose effects are manifest in our intelligent behaviour; in the other it refers to conscious mental acts involved in problem solving. In retrospect, this distinction anticipates the *dual-process* theories proposed by many contemporary cognitive and social psychologists, and the chapter shows how Dennett's distinction can be developed to provide an attractive version of dual-process theory. After introducing dual-process theories, the chapter reviews Dennett's remarks about thinking in *C&C* and shows how they suggest a reinterpretation of dual-process theory as a dual-*level* theory, grounded in the personal/sub-personal distinction also introduced in *C&C*. Later sections flesh out this theory, drawing on ideas from Dennett's later work, set out some of its attractions and implications, and show how it can be extended by combining it with a dual-attitude theory of belief also inspired by ideas in Dennett's work. The result is a picture of the human mind as a two-level structure, composed of a lower level of sub-personal informational states and processes and a higher, "virtual" level of personally constructed mental attitudes and operations.

4.1 Introduction

Content and Consciousness (hereafter, *C&C*)[1] outlined an elegant and powerful framework for thinking about the relation between mind and brain and about how science can inform our understanding of the mind. By locating everyday mentalistic

[1] References are to the second edition (Dennett 1986).

K. Frankish (✉)
The Open University, Milton Keynes, UK

The University of Crete, Heraklion, Greece
e-mail: k.frankish@gmail.com

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_4

explanations at the personal level of whole, environmentally embedded organisms, and related scientific explanations at the sub-personal level of internal informational states and processes, and by judicious reflection on the relations between these levels, Dennett showed how we can avoid the complementary errors of treating mental states as independent of the brain and of projecting mentalistic categories onto the brain. In this way, combining cognitivism with insights from logical behaviourism, we can halt the swinging of the philosophical pendulum between the "ontic bulge" of dualism (*C&C*, p. 5) and the confused or implausible identities posited by some brands of materialism, thereby freeing ourselves to focus on the truly fruitful question of how the brain can perform feats that warrant the ascription of thoughts and experiences to the organism that possesses it.

The main themes of the book are, of course, intentionality and experience – content and consciousness. Dennett has substantially expanded and revised his views on these topics over the years, though without abandoning the foundations laid down in *C&C*, and his views have been voluminously discussed in the associated literature. In the field of intentionality, the major lessons of *C&C* have been widely accepted – and there can be no higher praise than to say that claims that seemed radical 40-odd years ago now seem obvious. In the field of consciousness, the philosophical pendulum continues to swing, and Dennett continues to press the case for his position with clarity and wit. One can only hope that in another 40 years, these lessons too will seem obvious.

In this chapter I want to turn aside from these major themes to look at a relatively neglected part of *C&C* which I believe deserves to be better known by both philosophers and scientists. This is Dennett's discussion of thinking and reasoning in Chap. VIII. In this chapter Dennett distinguishes two senses of "thinking", parallel to the senses of "awareness" distinguished earlier in the book. In retrospect, this distinction anticipates contemporary "dual-process" theories of reasoning, and I shall show how Dennett's distinction might be developed and argue that it offers an attractive reinterpretation of the dual-process approach.

The chapter is structured as follows. Section 4.2 sets the scene by introducing dual-process theories in psychology. Section 4.3 reviews Dennett's remarks about thinking in *C&C* and shows how they suggest a version of dual-process theory conceived in terms of the personal/sub-personal distinction. Section 4.4 fleshes out this theory, drawing on ideas from Dennett's later work, and Sect. 4.5 outlines some attractions and implications of this version of dual-process theory. The final section shows how the proposed theory might be extended by combining it with a dual-attitude theory of belief, also inspired by ideas in Dennett's work.

4.2 Dual-Process Theories

In recent decades, researchers studying various aspects of human cognition have proposed dual-process theories. Such theories hold that there are two different processing mechanisms available for problem-solving tasks, usually labelled *Type 1*

and *Type 2*, which employ different procedures and may yield conflicting results.[2] Type 1 processes are typically characterized as fast, effortless, automatic, nonconscious, inflexible, heavily contextualized, and undemanding of working memory, and they are usually held to be responsible for biased and stereotypical responding on problem-solving tasks. Type 2 processes, by contrast, are typically described as slow, effortful, controlled, conscious, malleable, abstract, and demanding of working memory, and they are claimed to be the source of our capacity for normative responding in accordance with logical rules. Theories of this kind have been proposed, largely independently, by researchers on reasoning (e.g., Evans 1989, 2007; Evans and Over 1996; Sloman 1996; Stanovich 1999, 2011), decision making (e.g., Kahneman 2011; Kahneman and Frederick 2002; Reyna 2004), social cognition (e.g., Chaiken and Trope 1999; Smith and DeCoster 2000), and learning and memory (e.g., Dienes and Perner 1999; Reber 1993).[3]

In the field of reasoning and decision making, dual-process theories were originally proposed to explain conflicts between normative and biased responses on experimental tasks. However, the theories have subsequently been supported by a wide range of other evidence, including, (a), experimental manipulations (including explicit instruction) designed to shift the balance between the two types of processing (e.g., De Neys 2006; Roberts and Newton 2001), (b), psychometric studies showing that cognitive ability is differentially linked to performance on tasks where Type 2 thinking (which is demanding of resources) is required for production of the normative response (e.g., Stanovich 1999; Stanovich and West 2000), and, (c), neuroimaging studies indicating that responses associated with the different types of processing activate different brain regions (e.g., De Neys et al. 2008; Lieberman 2009; McClure et al. 2004). Theorists disagree about the relations between the two processes and about whether they operate in parallel or in sequence. A popular view is that Type 1 processes generate rapid default responses, which usually control behaviour but can, given sufficient resources, motivation, and ability, be intervened upon and replaced with more reflective responses generated by slower, Type 2 processes. Evans calls this view *default-interventionism* (Evans 2007).

Some dual-process theorists have taken a further step and proposed dual-*system* theories, according to which human cognition is composed of two multi-purpose reasoning systems, usually known as *System 1* and *System 2*, the former supporting Type 1 processes, the latter supporting Type 2 ones (e.g., Epstein 1994; Evans and Over 1996; Stanovich 1999, 2004). Dual-system theorists typically claim that System 1 is an evolutionarily old system, whose performance is unrelated to general intelligence, whereas System 2 is a more recent, uniquely human system, whose performance correlates with general intelligence.

Recently, however, some dual-process theorists have shunned the term "system", with its implications of unity, discreteness, and functional specialization, and

[2] As I use the term, dual-process theories contrast with dual-mode theories, which recognize the existence of two styles of reasoning but regard them as different modes of a single mechanism, or type of mechanism.

[3] For surveys of the literature, see Evans (2008), Frankish and Evans (2009), Frankish (2010).

reverted to talk of *types* of processing (e.g., Evans 2010; Stanovich 2011). It was always understood that System 1 was actually a suite of subsystems, including domain-specific modules, implicit learning mechanisms, emotional subsystems, and associations and responses learned to automaticity. And some theorists now argue that there are a variety of Type 2 systems as well, unified by their processing characteristics and shared use of working memory (Evans 2009). It has also been increasingly recognized that Type 2 processing requires supporting Type 1 processing of various kinds.[4] Theorists have also qualified their descriptions of the two types of process, stressing that many of the features commonly assigned to each should be treated as typical correlates rather than necessary, defining characteristics (Evans and Stanovich 2013). Thus on this view, Type 1 processes are often, but not invariably, contextualized, fast, and productive of biased responses, and Type 2 processes often, but not invariably, slow, abstract, and productive of normatively correct responses. At the same time, theorists have highlighted just one or two features as defining of Type 2, as opposed to Type 1, processing. For Evans it is use of working memory; for Stanovich it is reflective control and "cognitive decoupling", i.e., the capacity to entertain hypotheses and run mental simulations.

Although these qualifications soften the hard outlines of dual-system theory, they leave intact the core idea that there are two forms of cognition, one that is evolutionarily old, automatic, guided by instinct and habit, and independent of general intelligence, and another that is distinctively human, controlled, flexible, dependent on working memory, and linked to general intelligence. This broad picture is well supported. However, many issues remain, in particular about Type 2 processes. Type 2 processing seems capable of some prodigious intellectual feats. Indeed, it seems to occupy the role of something rather like a central executive, which can override instinctive, associative, and emotional responses with rational thoughts and decisions. Now, the positing of such an executive system is, of course, a move which Dennett opposes, as being both unexplanatory and neurologically implausible – a central theme of *Consciousness Explained* (Dennett 1991). Moreover, there are problems in explaining how Type 2 processing could have evolved. It is often claimed that Type 2 processing is evolutionarily recent and even unique to humans (e.g., Evans 2010), but this means it must have developed in a very short span of time on the evolutionary scale. A third issue concerns consciousness.[5] Type 2 thinking is usually characterized as being conscious, but there are reasons for doubting that conscious thought plays any *distinctive* role in guiding behaviour (as dual-process accounts assume it does). From a neural perspective,

[4] Evans, for example, stresses the role of preattentive Type 1 processes in supplying content to Type 2 processing and highlights the need for control processes that allocate resources to the two systems and resolve conflicts between them (Evans 2009).

[5] When I talk of *consciousness* in this chapter I mean *access consciousness* – roughly, availability to other central mental processes and to verbal report. The question is whether, in the case of thought, such access is (at least sometimes) associated with a different mechanism of behavioural control. I am not concerned with issues that arise specifically from the role of *phenomenal consciousness*, the putative subjective qualities of experience (for the distinction between access and phenomenal consciousness, see Block 1995).

consciousness seems to be a late-occurring event in the sequence from perceptual input to behavioural output, with conscious awareness of a decision lagging behind the neural processes that initiate it (e.g., Libet 2004). This suggests that the putative Type 2 processes may be merely side-effects of, or commentaries on, the non-conscious processes that do the real work in guiding behaviour (e.g., Wegner 2002) – a view which would be a single-process one.

I think there is a solution to these problems, which is compatible with the data and faithful to the spirit of dual-process theory. However, it requires a certain shift of perspective – a shift that can be motivated by looking at Dennett's views on thinking, beginning with those in *C&C*.

4.3 Dennett 1969 on Thinking

The discussion of thinking and reasoning in *C&C* comes in Chap. VIII, after Dennett has argued for a view of consciousness as availability to verbal report. On that view, content becomes conscious in virtue of becoming an input to the subject's speech centre – crossing the "awareness line". (In the terminology of the book, such contents are the objects of *awareness*$_1$. Contents that are effective in guiding behaviour but are not input to the speech centre are said to be objects of *awareness*$_2$; *C&C*, pp. 118–9). This is, of course, in stark opposition to the Cartesian view of consciousness as an internal arena where mental images are observed and mental operations performed, and Dennett's first task in discussing thinking and reasoning is to reject what he calls the "hammer and tongs" view, on which there are agents and objects in consciousness "[o]ne supposes that there are *conscious acts* of reasoning, acts of judgment and acts using concepts, and on the model of public acts we expect some organ, arm or tool to be *acting on* some object or some raw material – all this within the arena of consciousness" (*C&C*, p. 148).

In developing a better view, Dennett invokes the personal/sub-personal distinction. Reasoning, he maintains, is a personal activity – something done by persons, not by brains. But, most of the time, we have no awareness (i.e. awareness$_1$) of the processes that give rise to our conclusions and judgements, so this personal activity cannot be an operation or process *by which* a result is derived. Rather, it is more like the *reporting* of a result: "[a]s Ryle points out, such quasi-logical verbs as 'conclude', 'deduce', 'judge' and 'subsume' do not refer to processes at all, but are used in the presentation of results already arrived at" (op. cit., p. 149). Of course, there must *be* operations involved in the production of inferences and judgements, and these must be guided by stored information, but they will be of a sub-personal kind with no awareness$_1$. The hammer-and-tongs view results from projecting personal categories onto these sub-personal processes, and it involves a confusion of levels.

Dennett notes that this point may be obscured by the fact that we are often introspectively aware of *some* operations associated with problem solving:

> while engaged in problem solving we are aware$_1$ of a series of things prior to arriving at a conclusion, and we can often, on the basis of this awareness$_1$, divide our problem solving into a sequence of *operations* or *steps.* […] When one is asked how one figured out the answer, one can often give a list of steps, e.g., 'first I divided both sides by two, and then I saw that the left side was a prime …'. What one is doing when one reports these steps is by no means obvious. (*C&C*, pp. 150–1)

However, Dennett stresses, these operations, too, depend on sub-personal processes to fill in missing steps. We may be able to identify some personal-level operations in a problem-solving episode, but when we ask how we carried out *these* operations, we quickly run up against unanalysable personal activities, which must, nonetheless, be the product of complex sub-personal informational processing (*C&C*, pp. 151–2).

Now, the terms "thinking" and "reasoning" could refer either to sub-personal information processing whose effects are manifest in our intelligent behaviour, or to conscious mental acts involved in problem solving ("awareness$_1$ of an argument sequence leading to a conclusion" or "something like 'consciously reasoning with concepts'" as Dennett puts it; *C&C*, p. 155). Dennett notes that both usages have firm roots in everyday speech, and he suggests that the best course is to distinguish two senses of the words, parallel to the different senses of awareness distinguished earlier in the book. In the internal-processing sense animals can think, whereas in the conscious-acts sense they cannot (since Dennett identifies consciousness with awareness$_1$, which requires language). Moreover, thinking in the conscious-acts sense can be enthymematic, omitting important premises, whereas sub-personal processing cannot (*C&C*, p. 155–6). Sub-personal processes can fill in missing steps in our conscious reasoning, but those sub-personal processes themselves must draw, in some way or another, on all the information required to reach the conclusion. Dennett does not give names to these two types of thinking, but I shall call them *thinking$_1$* and *thinking$_2$*, the former being sub-personal information processing and the latter a process involving conscious mental operations of some kind. (This numbering unfortunately clashes with Dennett's numbering for awareness, where awareness$_1$ is the conscious form and awareness$_2$ the behaviourally manifest kind. However, it harmonizes better with the naming conventions in the psychological literature).

C&C distinguishes two senses of "thinking", then, but this is not yet a dual-process theory. The core feature of dual-process theories is the claim that there exist two different types of reasoning mechanism with different processing characteristics, as opposed to two different modes of a single mechanism. But, for all that has been said so far, thinking$_1$ and thinking$_2$ might be processes of the same type, differing only in that the latter happen to be conscious. That is, the episodes that are characteristic of thinking$_2$ might simply be episodes of thinking$_1$ that cross the awareness line by becoming inputs to the speech system. Such a view would not amount to a dual-process theory.

There are passages in *C&C* which might support this interpretation. Dennett stresses that in many cases much the same information processing must go on whichever type of thinking precedes an action. Whether we notice an apple and

consciously decide to eat it, or just pick up the apple and start munching, in either case our behaviour is guided by stored information about the edibility of apples, the ownership of this apple, the time to the next meal, and so on (*C&C*, p. 153). This does not count decisively against a dual-process view, however, either as an interpretation of *C&C* or as the best view of the situation. Dennett's main concern in this chapter is not to explore the nature of thinking$_2$, but to oppose the view that it is the only or core form of thinking and to argue that talk of thinking or reasoning is often simply an idealized intentional characterization of sub-personal information processing operations of which we have no conscious awareness. And while he does hold that the conscious events that are distinctive of thinking$_2$ are the product of thinking$_1$, this is not the same as saying that they are simply portions of thinking$_1$ that happen to be conscious.

Moreover, *C&C* contains hints of a different view. In discussing the role of awareness, Dennett notes that awareness$_1$ is associated with enhanced behavioural control: awareness$_1$ is a central component of attention, and attention improves control. However, he points out, awareness$_1$ *in itself* could not do this; reportability is not logically related to control. Rather, Dennett suggests, awareness$_1$ may be a contingent (and not invariable) by-product of a prior shift in control elsewhere in the system.

> There seem to be two levels from which we direct our behaviour. At the 'high' level (apparently in the cortex) we correlate information from a variety of sources, the behaviour controlled is versatile and changeable – and not particularly coordinated. Once under control, the behaviour is often made into a routine and the control is packed off into a more automatic and specialized system […] If 'paying attention' is a matter of dealing with the relevant parts of the environment at the high level, it might also *happen* to be a matter of bringing certain high-level signals across the awareness line, just because that is the way the brain is wired. (*C&C*, p. 124)

Dennett notes that such a contingent connection would be adaptive if it supported the practice of verbal instruction.

Combining these remarks with the distinction of types of thinking yields a dual-process picture, which posits two processes with different characteristics and mechanisms: thinking$_1$, which is non-conscious, effected by specialized subsystems, and supports fluid, unreflective behaviour, and thinking$_2$, which is typically conscious, is effected by higher-level mechanisms, and supports more flexible but less fluid behaviour. This outline picture harmonizes well with modern versions of dual-process theory.

C&C, then, offers an anticipation of modern dual-process theory. This in itself is interesting – further evidence of the book's far-sightedness and another instance of the independent emergence of dual-process views in different fields (Frankish and Evans 2009). However, the book also hints at something more. Dennett stresses that reasoning is a personal activity (*C&C*, p. 147–9). In the case of thinking$_1$ this means simply that verbs of thinking offer "fused" personal-level characterizations of cognitive accomplishments produced by sub-personal processes. However, thinking$_2$ may be a personal activity in a stronger sense. Dennett suggests that the *process itself* involves the performance of a sequence of personal-level actions; the conscious

operations or steps involved are intentional actions, like those involved in, say, talking or drawing, which are purposeful, require effort, and can be done well or badly. This view is hinted at earlier in Chap. VIII, where Dennett notes that there is a sense of "thinking" that connotes "purposeful and diligent reasoning, as in the sign on the office wall 'Think!'. […] In some way or other thinking in this sense, or reasoning, is a process, for it takes time, can leave us exhausted, go astray, be difficult, bog down" (*C&C*, p. 147). And it is a view that Dennett endorses explicitly in a later paper, where he writes that conscious propositional thinking "is a personal level activity, an intentional activity, something we *do*. […] It is not just something that happens in our bodies. When we think thoughts of this sort, we do, it seems, *manipulate* our thoughts, and it can be difficult or easy work" (Dennett 1998, p. 286).

This points to an alternative way of formulating dual-process theory, on which the core distinction is between reasoning processes that are wholly sub-personal (though the judgements and inferences they generate are ascribed to the person) and thinking that constitutively involves performing intentional, personal-level actions of some kind. (This is not to say that that is *all* that thinking of the second type involves; sub-personal reasoning processes may be involved in *generating* the actions in question. The claim is that what is distinctive of the second type of thinking is that it involves the performance of *some* intentional actions, of the appropriate kind).

I believe this is a fruitful approach, which harmonizes well with recent dual-process theories while offering solutions to some of the problems they face. The idea is only hinted at in *C&C*, and Dennett does not suggest what kind of personal activities might be involved in thinking$_2$. However, he returns to the subject, from a different perspective, in *Consciousness Explained* (Dennett 1991), and in the next section I shall show how the ideas there can be used to flesh out the proposed personal/sub-personal approach to dual-process theory.

4.4 Dennett 1991 on the Conscious Mind

In *Consciousness Explained* Dennett stresses the relative limitations of the human biological brain. Our brains, he points out, are little different from those of our ancestors 150,000 years ago. Fundamentally, they are collections of specialized but unintelligent subsystems, many innately specified, operating in parallel and competing for control of motor systems. They have, in addition, a high level of plasticity, conferring remarkable capacities for individual learning and adaptation, and they are promiscuous information gatherers across a range of sense modalities. But, considered as bare biological organs, human brains are largely driven by environmental stimuli and have little or no capacity for long-term planning or creative thought. The theoretical task, then, is to explain how such organs could support modern human minds, with their much enhanced powers: "[o]nto this substrate nervous system we now want to imagine building a more human mind, with something like a 'stream of consciousness' capable of sustaining the sophisticated sorts of

'trains of thought' on which human civilization apparently depends" (Dennett 1991, p. 189). Note that this is essentially the same question as that of how the brain came to support Type 2 processing (or System 2), with its capacity for decontextualized and hypothetical thought and higher-level, reflective behavioural control. And, indeed, what Dennett goes on to propose is in effect a version of dual-system theory.

Dennett argues that the conscious mind is too recent a development for it to be an innately specified biological system with a dedicated neural basis. Rather, he suggests, it is a softwired, or "virtual", system, which we create for ourselves by engaging in various culturally transmitted behaviours (memes or "good tricks"), which in effect reprogram our biological brains. The most important of these behaviours, he argues, is that of self-directed speech: producing, rehearsing, and rearranging sentences in overt or silent soliloquy (when audible, this is usually called "private speech", and when inaudible, "inner speech"). This stream of self-directed verbalization transforms the activity of the biological brain, causing its parallel, multi-track hardware to simulate the behaviour of a serial, single-track processor (Dennett 1991, Ch. 7).

Of course, the idea that thinking is a sort of inner monologue is not a new one. People often liken thinking to talking to oneself, and Ryle explored the idea in some depth, struggling with the problem of how to characterize the activity and its purpose in an illuminating way (Ryle 1979). However, Dennett offers a new slant on the role of self-directed speech. The fundamental idea is that such speech has a *self-stimulatory* effect. Self-generated utterances (questions, commands, reminders, and so on) are "heard" and processed like externally produced ones, and may evoke similar responses, with beneficial effects. Dennett highlights several aspects of this.

First, self-directed speech may promote information access among neural subsystems. A self-generated question may prompt a verbal reply, whose content will then be extracted by the speech comprehension system and made available to other neural subsystems, creating a "virtual wire" through which internally isolated subsystems can communicate. In this way, Dennett suggests, the channel of self-directed speech becomes an "open forum" where stored information can be accessed and applied to any problem (1991, pp. 194–7, 278).

Second, it may enhance behavioural control. Self-generated commands, exhortations, encouragements, and reminders can help to foster focused activity and prevent attention being captured by passing stimuli: "when a task is difficult or unpleasant, it requires 'concentration,' something 'we' accomplish with the help of much self-admonition and various other mnemonic tricks, rehearsals […], and other self-manipulations" (Dennett 1991, p. 277). Such manipulations, Dennett suggests, achieve their effects by co-ordinating the activities of specialist subsystems: they serve to " adjudicate disputes, smooth out transitions between regimes, and prevent untimely *coups d'etat* by marshaling the 'right' forces" (ibid.).

Third, Dennett stresses that self-directed speech facilitates hypothetical thinking and long-term planning ("producing future"). The idea is that self-generated scenarios or proposals may provoke thoughts of their likely consequences, allowing one to assess courses of action in advance of performing them. Saying to oneself,

"What if I did this?" may stimulate thoughts and images of the consequences of doing it, generating positive or negative reactions and so allowing one to evaluate the proposed the action.[6] Dennett suggests that self-directed speech can also assist planning by reinforcing memory. Commenting on one's actions can make it easier to keep track of the progress one has made and to recall the strategies one has used and their success, thus helping one to choose wisely in future (1991, p. 278).[7]

Once the trick of self-directed speech had been discovered, Dennett argues, it would have been refined by suppression of overt vocalization and disseminated and elaborated through processes of cultural evolution, and a disposition to master it might have been coded into the human genome, thanks to the Baldwin effect.[8] Other forms of internalized self-stimulation also emerged and spread, including the manipulation of visual imagery ("diagramming to oneself") as a private substitute for diagram drawing (op. cit., p. 275).[9] As a result, Dennett claims, we have become disposed to develop habits of regular inner speech and other forms of self-stimulation, thereby artificially creating a new level of cognitive activity, which is both serial and heavily language-involving. Dennett calls this softwired system the *Joycean machine* (after James Joyce's 1922 novel *Ulysses*, which records its characters' inner monologues).

This, then, offers an account of the intentional actions involved in thinking$_2$: they are self-stimulations, typically abbreviated, internalized speech acts. Does this mean that we should, after all, accept a version of the hammer-and-tongs view of thinking derided in *C&C*, on which there are conscious acts of reasoning, performed upon propositional objects? Only in a very weak sense. On the proposed view there are acts and objects in reasoning, but they are acts of the person, not of an inner,

[6] Dennett gives an example using private diagram drawing, which he claims can also be used for self-stimulation (1991, p. 197, pp. 220–1).

[7] Carruthers has proposed a similar account of how self-directed speech supports hypothetical thinking, developed within the context of a massively modular view of the mind (Carruthers 2006, 2009).

[8] The idea is that once an individual has discovered a useful behaviour, such as the knack of making a certain tool, selectional pressure will arise for other members of its community to acquire it too. Those who find it easy to learn the behaviour will be selected for over those who find it hard, and, over time, individuals who are predisposed to learn it will come to predominate in the community (e.g., Dennett 1991, pp. 184–7, 2003).

[9] Note that this assumes that we can intentionally generate sensory imagery. This might involve the mental rehearsal of action, as proposed by Carruthers (Carruthers 2006, 2009). The idea is that when an action schema is activated, an internal efference copy of it is created, which is used to create a "forward model" of the action. This then generates proprioceptive and other sensory representations of the movements involved, which are used to guide the execution of the action and anticipate its consequences. In mental rehearsal, Carruthers argues, action schemata are activated offline, with the muscle commands suppressed but the efference copies still issued. The sensory images produced are then received by input systems (audition, vision, speech comprehension, etc.), and the information they carry globally broadcast to modular subsystems. Where the rehearsed action is an utterance, auditory images (inner speech) are produced and interpreted, and their contents broadcast.

quasi-Cartesian agent, and their objects are sentences, or images of sentences, not wordless propositions.

It is worth stressing that to ascribe these acts to the person is not to ascribe them to something *additional* to the set of sub-personal cognitive mechanisms and other biological subsystems that collectively compose the person. The person is not some extra component or feature with new causal powers, and personal activities are performed in virtue of sub-personal processes (note how Dennett puts "we" in inverted commas in the quote above). Explanations in terms of the person derive from adopting a certain interpretive stance toward a biological entity – viewing it as a unified organism, embedded in, and generally well-adapted to, its environment, and with global behavioural dispositions and susceptibilities. It is nonetheless important to make the personal/sub-personal distinction, since personal properties and activities may involve the coordinated activity of many separate sub-personal systems, whose importance and explanatory role is visible only from the personal perspective. Thus, personal-level, self-stimulatory reasoning involves many subsystems that are not involved in wholly sub-personal reasoning, including motor systems, working memory, and sensory systems.[10]

It may be objected that Dennett's proposal does not achieve its aim of explaining higher-level thought without positing something like a central executive. Intentional actions are motivated by beliefs and desires – desires to achieve ends and beliefs about the means to achieve them. And this might suggest that before engaging in self-stimulation, we must have determined what cognitive and behavioural effects we want to achieve and worked out how sub-personal systems need to be stimulated in order to achieve them. But if so, then all the real problem-solving work would already have been done, presumably by some executive system, and the stimulation would be merely a mechanism for implementing its conclusions – and a clumsy and inefficient one at that (why not pass on the executive's commands directly through internal channels?). Dennett anticipates this objection, of course. He argues that speech production need not be the product of specific intentions formulated by an executive system (a "Conceptualizer" or "Central Meaner"), which figures out what needs to be said in advance. Rather, he suggests, sophisticated speech acts might be generated through a process of quasi-evolutionary competition between numerous

[10] It may be asked whether it is legitimate, within the framework of *C&C*, to talk of personal-level actions *causally affecting* sub-personal information processing. After all, Dennett repeatedly cautions against confusing the levels throughout the book. It is true that strict causal explanations of sub-personal events will be framed wholly in sub-personal terms, but we can talk loosely of token personal events having sub-personal effects, provided the events in question are identical with sub-personal ones. And although Dennett denies that *some* personal events (pains, for example) can be identified with sub-personal ones (*C&C*, p. 94), he does not issue a blanket ban on personal-sub-personal identifications, and suggests that we proceed on a case-by-case basis (pp. 16–18, 96). (At the extreme we can treat the personal descriptions as fused and identify them with descriptions of global physical state; Dennett 1987, p. 57). In the case of imagistic self-stimulations it is plausible to think that at least rough identifications can be made with sequences of sub-personal events, perhaps involving the offline activation of motor schemata, and causal explanations mentioning them should be understood as shorthand for more rigorous but less perspicuous explanations couched in such terms.

unintelligent micro-systems (a "pandemonium" of "word demons"), vying to produce utterances of varying degrees of sophistication and appropriateness (Dennett 1991, Ch. 8). We might ascribe a sophisticated communicative intention to the *speaker*, but it need not correspond to any prior sub-personal command. Similarly, self-directed speech acts might be generated pandemonium-style, without antecedent calculation of their structure or likely effects. It is true that, if they are to count as intentional, self-stimulations must be susceptible to some intentional characterization, but this need not be in terms of desires for specific cognitive and behavioural effects and beliefs about how to achieve them. The motivating states might simply be a desire to solve some problem and the instrumental belief that doing *this* (uttering the words that spring to one's lips) may help.

But could pandemonium processes generate the subtle self-stimulations required to support executive control, abstract problem solving, and hypothetical thinking? Where does the intelligence in these acts come from? There are several points that might be made here. First many self-stimulations, verbal and otherwise, are *not* particularly intelligent. Much self-directed speech consists of comments on what is happening, chance associations, whimsy, free-wheeling speculations, and so on (just like the monologues in Joyce's *Ulysses*). Useful queries, exhortations, and ideas might be simply chance products of this continual stream of commentary.[11] Second, acts of self-stimulation often form part of a sequence of such acts (trains of thought). Self-generated speech and other imagery may not only stimulate cognitive and affective responses, but also trigger further acts of self-stimulation, shaped by those responses. For example, imagining a course of action may provoke images of the action's likely consequences, which may then suggest images of further actions, and so on. In this way, cycles of self-stimulation may arise, taking unanticipated and creative directions (Carruthers 2006, Ch. 5). Third, self-stimulation may be guided by knowledge imparted by culture. Cultural processes may disseminate, not only the trick of self-stimulation itself, but specific applications of it to particular problems. Think, for example, of mnemonic rhymes, like that for the number of days in the months. If we know the rhyme, we can literally *tell* ourselves how many days there are in each month, even if we cannot recall the information directly. There are countless other problem-solving routines we can learn, involving inner speech or inner diagramming, which embody logical or mathematical principles or heuristics of various kinds. More broadly, we can also learn ways of enhancing self-stimulation through developing habits of self-questioning, self-commentary, and so on – habits sometimes taught under the heading of "metacognition". As Dennett stresses, the distinctive power of the Joycean mind is due far more to cultural programming than to the underlying biological hardware.

Combining these elements from *C&C* and *Consciousness Explained*, we have a dual-process theory which distinguishes sub-personal informational processes (thinking$_1$) and processes involving personal-level intentional self-stimulation

[11] Note, too, that self-stimulation may not always be beneficial. There can be negative thinking as well as positive, and habits of harmful self-stimulation may contribute to some psychopathologies, such as anxiety, depression, and obsessive-compulsive disorder.

(thinking$_2$). In the next section I shall compare this approach with standard dual-process theories, indicate some of its theoretical attractions, and suggest how it might be experimentally evaluated.

4.5 Connections, Attractions, and Implications

In outline at least, Dennett's dual-process theory harmonizes well with other recent dual-process theories. The features of sub-personal information processing and personal self-stimulation coincide with those usually ascribed to Type 1 and Type 2 processes. Typically, sub-personal processes are fast, effortless, automatic, non-conscious, and inflexible, whereas acts of intentional self-stimulation are slow, effortful, controlled, conscious, and malleable. Sub-personal processes are also likely to display far less individual and cultural variation than processes of personal self-stimulation.

Moreover, identifying Type 2 processes with acts of self-stimulation explains *why* Type 2 processes have the features they do. They are slow because they employ serial channels designed for speech production and comprehension or for other forms of mental rehearsal (Dennett 1991, p. 197). They are controlled and effortful because they are intentional actions that demand attentional resources. They are conscious because (as Dennett notes) they are perceived just like external stimuli: inner speech is processed by the auditory system, inner diagramming by the visual system, and so on (Dennett 1991, pp. 225–6). They are malleable because, like other intentional actions, they are responsive to beliefs about how they should be conducted – that is, about what problem-solving routines are normatively warranted (Carruthers 2009). They exhibit high individual variation because individuals differ in their attentional resources and self-regulatory dispositions, and they exhibit high cultural variation because different cultures inculcate different self-stimulatory habits and different problem-solving strategies. At the same time, however, these features are not essential or unqualified. For example, some well-practised self-stimulatory routines could be relatively swift and effortless. This again is in line with current thinking by dual-process theorists (e.g., Evans and Stanovich 2013).

More importantly, self-stimulatory processes exhibit the two core features of Type 2 processes identified by Evans and Stanovich: use of working memory and cognitive decoupling. Self-stimulation draws on working memory because it involves attending to and manipulating sensory imagery. (Indeed, it is arguable that working memory just *is* the set of resources involved in manipulating sensory imagery in the service of self-stimulation; Carruthers 2006). And self-stimulation supports cognitive decoupling since inner speech and other self-generated sensory images can represent non-actual scenarios.

Given these similarities, evidence that supports standard dual-process theories, with the features mentioned, also supports the Dennettian version. And, since the sub-personal/personal distinction explains the other distinctions, including the supposedly core ones, it has a claim to be the truly fundamental distinction, from which all the others follow.

It may be objected that this is not a genuine dual-process theory: although Type 2 thinking involves mechanisms of self-stimulation not involved in Type 1 thinking, all the real *inferential* work is done by the sub-personal reasoning processes that first generate and then respond to sensory imagery. Now, it is true that, on the proposed view, Type 2 reasoning constitutively involves passages of Type 1 reasoning (together with the activity of other motor systems, working memory, and sensory systems), but that is not to say that it is simply a *mode* of Type 1 thinking. There are two points to make, concerning the form and content of Type 2 thinking. Concerning form, on the proposed view, dual-process theory picks out different levels of functional organization, one partially realized in the other, and processes at different levels may have very different formal characteristics. Just as parallel connectionist processes can be implemented on a suitably programmed serial computer, so (the claim is) slow, controlled, and serial Type 2 processes are (partially) implemented in fast, automatic, and parallel Type 1 processes. The facts about implementation do not impugn the reality of the processes implemented. Second, concerning content, the reasoning at the two levels will be directed to subtly different problems. Personal-level processes will be directed to some real-world problem, whereas the supporting sub-personal activity will be devoted to the problem of producing and evaluating sensory imagery relevant to that problem. For example, suppose I have to work out how to fit various differently sized objects inside a box, and that I do this by imagining various possible arrangements to find one that works. Here *I* am thinking about the objects and the box, whereas the underlying sub-personal processes are concerned with what sensory images to produce, the interpretation of these images and the scenarios they represent, and what further images to produce. These processes become relevant to the box problem only when viewed as part of an extended process of problem solving involving the box and my purposes concerning it – a perspective that requires a shift to the personal level.[12] (Of course, this is not to say that sub-personal reasoning processes never engage directly with real-world problems. Much of the time they do just that, guiding behaviour without the occurrence of any self-stimulatory processes. They have the character described only when they are supporting Type 2 reasoning.) I conclude that Dennett's approach qualifies as a genuine form of dual-process theory, albeit an unorthodox one. The approach has, moreover, several theoretical attractions, as I shall now explain.

First, as we have seen, the approach offers an explanation of how Type 2 processing could be implemented without a dedicated executive system. On this view, executive functions are performed by temporary coalitions of specialist subsystems, formed under the influence self-stimulatory habits:

> In our brains there is a cobbled-together collection of specialist brain circuits, which, thanks to a family of habits inculcated partly by culture and partly by individual self-exploration, conspire together to produce a more or less orderly, more or less effective, more or less

[12] It is true (as mentioned earlier) that the results of thinking$_1$ (the decisions, conclusions, and actions) and the beliefs and desires that explain these results, are ascribed to the person too. Thus in a sense there are two levels of personal activity here, and I solve the box problem *by* solving (in a constitutive sense) the problem of how to stimulate myself in relevant ways. But the inferential operations involved in solving the latter problem are wholly sub-personal.

> well-designed virtual machine, the *Joycean machine*. By yoking these independently evolved specialist organs together in common cause, and thereby giving their union vastly enhanced powers, this virtual machine, this software of the brain, performs a sort of internal political miracle: It creates a *virtual captain* of the crew, without elevating any one of them to long-term dictatorial power. (Dennett 1991, p. 228)

Putting it another way, on this view, the executive system is nothing smaller than the person – the whole system acting upon itself through self-stimulation.

Second, and relatedly, Dennett's approach explains how Type 2 thinking could have evolved in such a relatively short space of time. The key biological developments were that of a language system, a capacity for the mental rehearsal of action, and supporting working memory resources. It is plausible to think that all of these evolved independently and for other purposes, and that Type 2 thinking involved their collective exaptation for cognitive purposes (Carruthers 2006). The other developments required were cultural, not biological, and included the invention, dissemination, and refinement of personal reasoning strategies, perhaps reinforced by the emergence of some innate dispositions to master such strategies, fostered by the Baldwin effect. These developments could have been extremely rapid on the evolutionary scale.

Third, the approach explains how there can be a distinct role for conscious thought, even if consciousness is a late event in neural processing. The key point is that the last event in one cycle of processing can become the first in a new cycle, via the loop of self-stimulation. To forget this, Dennett remarks, "is like forgetting that the end product of apple trees is not apples – it's more apple trees" (1991, p. 255). Indeed, understood as imagistic self-stimulations, conscious thoughts become cognitively effective in virtue of the very same fact that makes them conscious: namely that they are received and processed by sensory input systems.

Thus, Dennett's version of dual-process theory harmonizes well with existing ones and has some distinct advantages as well. However, the theory also differs in important ways from standard dual-process accounts, in particular in its implications for the way Type 2 thinking is implemented in the brain, and I shall conclude this section by highlighting some of these differences.

The first concerns the relation between the neural bases for Type 1 and Type 2 thinking. Psychologists tend to think of the two processes as associated with distinct neural structures. However, on Dennett's view, this will not be the case. On this view, Type 2 thinking does not engage distinct neural mechanisms but involves the exploitation and coordination of Type 1 mechanisms. As Dennett puts it, the installation of the Joycean machine "is determined by myriad microsettings in the plasticity of the brain, which means that its functionally important features are very likely to be invisible to neuroanatomical scrutiny in spite of the extreme salience of the effects" (Dennett 1991, p. 219). It is true that Type 2 thinking will engage some *additional* mechanisms, such as working memory, that are not involved in Type 1 processing, and the theory thus predicts that there will be some salient differences in the patterns of neural activity associated with each type of thinking (as the evidence indicates there are), but many processing resources will be shared between them.

A second consequence concerns language. It is often claimed that Type 2 thinking is linked to language, in that it is directly responsive to verbal instruction in a way that Type 1 thinking is not. However, if Dennett is right, there will be a much stronger link between Type 2 thinking and language. For the resources of the language system, including both language production and comprehension, will be constitutively involved in Type 2 thinking. As Dennett puts it, "a large portion – perhaps even the lion's share – of the activity that takes place in adult human brains is involved in a sort of word processing: speech production and comprehension, and the serial rehearsal and rearrangement of linguistic items, or better, their neural surrogates" (1991, p. 225). Of course, Type 2 thinking may involve non-linguistic forms of self-stimulation, too, such as inner diagramming, but the enormous representational powers of language will make it the dominant form. Note, too, that though Dennett talks of speech here, it would be better to say *language*. I assume that Joycean processes could be implemented in sign language instead of speech, either with overt signing, or covertly, using proprioceptive or visual imagery.

Further consequences follow from the status of Type 2 processes as intentional actions. First, as *actions*, they will involve the activation of brain regions associated with behavioural control, such as the motor and premotor cortex. Second, as *problem-solving* actions, they will draw on metarepresentational resources. As noted earlier, self-stimulatory activities will typically be motivated by a desire to solve some problem and instrumental beliefs about the strategies that may work. That is, engaging in Type 2 thinking involves thinking not only about the first-order problem one faces but also about the meta-problem of how to solve this problem, and it will therefore draw on metarepresentational and metacognitive resources. This meta-level thinking will, of course, usually be of the sub-personal, Type 1 kind.

The consequences mentioned suggest ways in which Dennett's version of dual-process theory may be experimentally tested, using techniques such as deficit studies, neuroimaging, and dual-task methodologies (in which a subject is required to perform two tasks simultaneously, in order to determine if they share processing responses). The fate of the theory will ultimately depend on such investigations. For the present, however, I shall conclude by highlighting a further theoretical attraction of Dennett's approach, which lies in the way it can be extended.

4.6 From Dual Processes to Dual Attitudes

Dual-process theories are often combined with what I shall call *dual-attitude* theories, according to which each type of processing has its own memory system, with a distinct set of propositional attitudes (e.g., Reber 1993). This view is supported by social-psychological work on persuasion and attitude change, which has led several theorists to distinguish two memory systems: an implicit system, which is non-conscious, automatic, fast-access, and slow-learning, and an explicit system, which is conscious, effortful, slow-access, and fast-learning (e.g., Wilson et al. 2000; Smith and DeCoster 2000; Smith and Collins 2009). Again, Dennett's work

contains the seeds of an alternative approach to dual-attitude theory, also rooted in the sub-personal/personal distinction first introduced in *C&C*. There is no space here to develop the approach in detail, but I shall sketch the outline (for more details, see Frankish 2004).

The key ideas appear in a chapter in *Brainstorms* (Dennett 1978, pp. 300–9), which looks at the operation of changing one's mind. Here Dennett endorses a suggestion by Ronald de Sousa (de Sousa 1971) that we have two levels of belief: graded, nonverbal belief, which is common to humans and animals, and binary, verbalized belief, which results from assenting to a natural language sentence. De Sousa likens the act of assent to a "*bet on truth alone*, solely determined by epistemic desirabilities" (quoted in Dennett 1978, p. 304). The product of such an epistemic bet, Dennett suggests, is not so much a belief as a state of commitment or ownership. Assenting to a sentence involves metaphorically putting it in a box marked "True" and committing oneself to asserting it in appropriate contexts. Dennett calls these commitments "opinions" and notes that a person's opinions may diverge from their nonverbal beliefs, as manifest in their behaviour. To make up, or change, one's mind about something, Dennett proposes, is to form or revise an opinion. (He adds, however, that not all opinions are the product of deliberation; some sentences are such sure bets that we add them to our collection of opinions without thinking).

Dennett occasionally invokes the belief/opinion distinction in his later work, but he does not build on it and does not connect it with his account of the Joycean machine. However, it is natural to make such a connection. The key move is to suppose that the commitment involved in opinion formation extends not only to asserting the endorsed sentence in public, but also to holding it true in one's private self-stimulatory activities. This might involve telling oneself that it is true, taking it as a premise when constructing explicit arguments, rejecting sentences that conflict with it, and so on. (We might say that if conscious reasoning is an exploration of a theoretical landscape – a metaphor used by Ryle (2009) – then opinions are signposts we erect along the way). On this view, a person's opinions will shape the course of their self-stimulatory activities and the cognitive and behavioural effects that result, functioning very much as beliefs are supposed to do. Indeed, on this view, it becomes attractive to redescribe the belief/opinion distinction as one between types of belief, Type 1 and Type 2, the former associated with Type 1 reasoning and the latter with Type 2. We might also identify a parallel Type 2 form of desire, which involves committing oneself to taking a sentence as a statement of a goal and treating it as a fixed point in our self-stimulatory activities.

This, then, offers a Dennettian dual-attitude view to complement the dual-process one. The theory retains the overall characteristics of standard accounts. Type 1 beliefs are formed slowly, through exposure to environmental regularities, but they influence behaviour rapidly and without conscious thought. Type 2 beliefs, on the other hand, can be formed rapidly through one-off acts of assent, but they influence action only through slow and effortful self-stimulation. Moreover, they are activated only when the agent is engaged in self-stimulation, or prompted to engage in it. In contexts where there are no prompts to self-stimulation they remain inert.

However, like the companion dual-process view, this view differs from standard ones in that it is rooted in the sub-personal/personal distinction. Although both types of belief are ascribed to persons, the processes involved in their formation and behavioural manifestation are located at different levels. Type 1 beliefs are formed by sub-personal processes, and they manifest themselves directly in behaviour without intervening personal activity. Type 2 beliefs, on the other hand, are formed through personal acts of assent, and they influence behaviour via personal-level self-stimulation. As with the companion account, this has consequences for the neural basis of the higher-level states. On the proposed view, Type 2 beliefs are commitments, and commitments can be analysed as complexes of beliefs and desires. Simplifying somewhat, to be committed to doing something is to *believe* that one has committed oneself to doing it and to desire to honour one's commitments. So, to have the Type 2 belief that *p* is to believe that one has committed oneself to holding true a sentence with content *p* and to want to honour this commitment.[13] (A similar analysis holds for Type 2 desire.) These constituting attitudes need not themselves be Type 2 ones, of course, and typically will not be (and if they are, the attitudes constituting *those* attitudes will surely not be). Thus, on this view, Type 2 attitudes are ultimately constituted by a set of metarepresentational Type 1 attitudes. The upshot is that, like the Joycean machine, the Type 2 memory system is a soft-wired, virtual one, realized in Type 1 states. Again, this view has attractions from an evolutionary perspective and means that Type 2 attitudes will not have a separate neural basis.

Dennett's dual-process theory thus naturally extends to give a picture of the human mind as a two-level structure composed of a lower level of sub-personal informational states and processes and a higher, "virtual" level of personally constructed mental attitudes and operations, which is constitutively dependent on the former. This is, I suggest, another reason to prefer it.

4.7 Conclusion

One of Dennett's aims in *C&C* was, I take it, to correct the error of philosophers who overlooked the existence of sub-personal reasoning processes or mischaracterized them in personal terms. But if the account sketched here is correct, then some contemporary psychologists are making the complementary error of overlooking personal reasoning processes or mischaracterizing them as sub-personal. Highlighting this possible error is one of the many salutary tasks *C&C* still performs in the fifth decade after its publication.[14]

[13] This is argued in detail in Frankish (2004).

[14] I thank the editors of this volume for their helpful comments on an earlier draft of this chapter. Thanks are also due to Jonathan Evans, Eileen Frankish, Liz Irvine, and Maria Kasmirli for their comments and advice.

References

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–287.

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought*. Oxford: Oxford University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 109–127). Oxford: Oxford University Press. (See https://www.plymouth.ac.uk/staff/jonathan-evans).

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17*(5), 428–433.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science, 19*(5), 483–489.

De Sousa, R. B. (1971). How to give a piece of your mind: Or, the logic of belief and assent. *The Review of Metaphysics, 25*(1), 52–79.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Montgomery: Bradford Books.

Dennett, D.C. (1986). *Content and consciousness* (2nd ed.). London: Routledge and Kegan Paul (First edition 1969).

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown and Co.

Dennett, D. C. (1998). Reflections on language and mind. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 284–294). Cambridge: Cambridge University Press.

Dennett, D. C. (2003). The Baldwin effect: A crane, not a skyhook. In B. H. Weber & D. J. Depew (Eds.), *Evolution and learning: The Baldwin effect reconsidered* (pp. 69–79). Cambridge, MA: MIT Press.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*(5), 735–808.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *The American Psychologist, 49*(8), 709–724.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove: Lawrence Erlbaum Associates.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove: Psychology Press.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). Oxford: Oxford University Press.

Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford: Oxford University Press.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove: Psychology Press.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*(10), 914–926.

Frankish, K., & Evans, J. St. B.T. (2009). The duality of mind: An historical perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–29). Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge: Cambridge University Press.
Libet, B. (2004). *Mind time: The temporal factor in consciousness*. Cambridge, MA: Harvard University Press.
Lieberman, M. D. (2009). What zombies can't do: A social cognitive neuroscience approach to the irreducibility of reflective consciousness. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 293–316). Oxford: Oxford University Press.
McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306*(5695), 503–507.
Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.
Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science, 13*(2), 60–66.
Roberts, M. J., & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 54*(4), 1031–1048.
Ryle, G. (1979). *On thinking*. Oxford: Blackwell.
Ryle, G. (2009). The thinking of thoughts: What is "Le Penseur" doing? In G. Ryle (Ed.), *Collected papers colume 2: Collected essays 1929–1968* (pp. 494–510). Abingdon: Routledge.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.
Smith, E. R., & Collins, E. C. (2009). Dual-process models: A social psychological perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 197–216). Oxford: Oxford University Press.
Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108–131.
Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah: Lawrence Erlbaum Associates.
Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.
Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–726.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101–126.

Chapter 5
The Rationality Assumption

Richard Dub

Abstract Dennett has long maintained that one of the keystones of Intentional Systems Theory is an assumption of rationality. To deploy the Intentional Stance is to presume from the outset that the target of interpretation is rational. This paper examines the history of rationality constraints on mental state ascription. I argue that the reasons that Dennett and his philosophical brethren present for positing rationality constraints are not convincing. If humans are found to be rational, this will not be because a presumption of rationality must be built into the deployment of the Intentional Stance. It will be an empirical finding. Rationality will be an outcome of mental state ascription rather than a condition on ascription.

5.1 Forefathers

Daniel Dennett studied under Quine at Harvard and under Ryle at Oxford. It is only moderately procrustean to say that Intentional Systems Theory is what you get by stirring together Quinean and Rylean metaphysics of mind. Quine provided tough-minded naturalism and an emphasis on the holistic, indeterminate, and irreducible nature of intentional language; Ryle provided a sensitivity to ordinary language that resisted eliminating mental talk as a dispensable dramatic idiom. Dennett's signature ingenuity was the alchemical spark needed to catalyze the reaction between the two.

Nowhere are Dennett's twin influences as keenly felt as in his first book. *Content and Consciousness* offers a germinal version of the Intentional Systems Theory that Dennett still maintains to this day. The debt to his philosophical forefathers in the book is explicit, and this makes it an especially fruitful place to turn to when attempting to fit the Intentional Stance within a historical tradition. In this chapter, I'll be exploring the history of one of the more controversial features of Intentional Systems Theory: its adherence to *a rationality assumption on belief ascription*. According to Dennett (both then and now), to apply the Intentional Stance – that is, to interpret an individual as having a mind – involves an assumption that the individual is rational.

R. Dub (✉)
University of Geneva, Geneva, Switzerland
e-mail: richard.dub@gmail.com

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_5

Many philosophers bristle at the suggestion. Without some fancy footwork, the claim that believers are necessarily rational simply looks empirically false. One occasionally gets the sense that some philosophers think the rationality requirement is not just wrong: they think it is absurd. Because of this, the force of their arguments comes down to their ability to convey goggle-eyed incredulity through text.[1] In addition to the "obvious" irrationalities we experience in ourselves and in others, there is scarcely an end to findings in psychology and behavioral economics purport to demonstrate various ways in which we are all exceedingly irrational. For instance, the work of Kahneman and Tversky is often presented as evidence for a natural human tendency to make various errors in probabilistic or conditional reasoning (Kahneman et al. 1982; Thagard and Nisbett 1983; Stich 1985; Cherniak 1986).

It is perfectly legitimate to argue against the rationality assumption by offering apparent counterexamples, but the method does not get to the heart of the matter. In what follows, I challenge the rationality assumption by challenging Dennett's need for such an assumption in the first place. Why does Dennett argue for a rationality assumption at all? What functions is it meant to serve? If these functions are legitimate, can they be served by other means? There are two similar but distinct arguments for the need for a rationality requirement in Intentional Systems Theory, each bequeathed to Dennett by his philosophical forebears. One of these arguments comes from Dennett's Quinean heritage; the other comes from his Rylean side. I'll develop these lines of argument, and show that neither is successful.

Thus, the main goal of this paper is to diagnose and reject Dennett's stated need for a rationality assumption. However, this leaves us with a new problem. What is the upshot if the arguments for the rationality assumption are unsuccessful? Should we drop the assumption? What would happen were it dropped? I'll argue (as a secondary thesis) that, in the end, not very much would change. A version of Intentional Systems Theory without a rationality assumption won't necessarily end up rendering the verdict that our neighbors are irrational. In fact, it might still well have us ascribe largely rational beliefs.[2] On this version of Intentional Systems Theory, the rationality of our neighbors (if they are indeed rational) will be an empirical finding rather than something to be settled before empirical investigation has begun.

5.2 The Quinean Lineage

Dennett is not the only figure who has argued for rationality constraints; Donald Davidson (1982) and David Lewis (1974) also include rationality constraints in their theories of mind in the form of "principles of charity." It's not surprising that there should be theoretical affinities between these three. All are interpretivists,

[1] A parody argument: "It's simply irrational to conclude that people are rational! Therefore, Dennett's theory is self-refuting."

[2] The extent to which people are actually rational or irrational is something that I will remain agnostic about for the purposes of this piece.

holding that interpretation is an important feature in the assignment of mental states. But more importantly, all are students of Quine. This gives us reason to analyze them together. (When studying an organism, if you don't know what a particular anatomical structure is for, it is sensible to look at homologous structures in the organism's ancestors and cousins. Likewise, it makes sense to look at the development of rationality requirements within this philosophical clade to see what similarities and differences we can tease out.) An ancestral form of the principle of charity can be found in Quine's *Word and Object* (1960b), so it makes sense as a starting point for our investigation.

Quine's principle of charity first appears during a discussion of radical translation. Quine famously argued that translation between languages would always be beset with indeterminacy. However, the existence of multiple contending translations does not mean that anything goes and that no translation is better than any other. Quine argued that in addition to respecting stimulus meanings, we ought to abide by certain maxims of translation which would have us prefer certain translation manuals to others. The principle of charity is one such maxim: it would have us rule out translations resulting in logical silliness. Take Quine's field linguist, charged with translating a language he has never heard before. He notices that speakers always assent to utterances of the form ⌜q ka bu q⌝. This counts as evidence against translating 'ka' as 'and' and 'bu' as 'not'. Such a translation would have the speaker assenting to contradictions, and so imputes unacceptable silliness. The principle of charity is what motivates Quine's (1960a) famous declaration that "prelogicality is a myth of bad translators."

Those following in Quine's footsteps took the principle of charity to be inculpated in projects wider than just linguistic translation. For Davidson and Lewis, the principle of charity is a constraint that preserves rationality during radical interpretation. Radical interpretation is unlike radical translation in that it is not purely linguistic; it also ascribes mental states to an agent. The principle of charity here is much the same. Quine introduced a principle of charity on radical translation to rule out translation manuals that would impute logical silliness; the reason for introducing a rationality constraint on radical interpretation is to pare down on an otherwise unbridled indeterminacy that would plague mental state ascription.

What is the source of such unbridled indeterminacy? In presenting his argument for the rationality assumption in *Content and Consciousness*, Dennett includes a particular argument of Quine's. It is worth quoting Dennett at length:

> Quine and Chisholm also present arguments about believing and intending, of which the central point is that efforts to provide behavioural analyses of these two phenomena are doomed by a vicious circle of implications. Take, for example, the belief that it is raining. What behavior would clinch it that A believes it is raining? No matter what is suggested, it will turn out that this is a clincher demonstrating that A believes it is raining *only* if we assume that A has some particular purpose or intentions. […] A's finding a tree or roof to stand under is no more evidence, for it depends on A's intending to stay dry. If ascription of belief always depends on an assumed ascription of intention, the converse holds as well. A's intention to stay dry is not behaviorally demonstrated by his cowering under the tree except on the assumption that he believes it is raining, that he believes that he would get wet if he

> did not stay under cover, and so forth. A survey of the other Intentional and mongrel Intentional idioms shows that the use of any one of them has implications about beliefs and intentions, so the circle that prevents a behavioural paraphrase of belief and intention sentences infects the whole realm of the Intentional (Dennett 1969, 31–2).

Dennett goes on to discuss how this argument establishes the holistic nature of mentalistic vocabulary, and therefore its irreducibility to a purely extensional language. But he also takes this section to establish that "intentional explanations presuppose the appropriateness of sequences they purport to explain." That is, this section is also taken to establish the rationality of the actor.

How does it do so? If we see A standing under a tree, we could interpret him as having a desire to stay dry, a belief that he'll stay dry if he stands beneath the branches, and an intention to do so. Or, we could interpret A as *wanting* to get wet, and believing that he'll get wet by going into the rain, but *irrationally* deciding to stay under the tree. We could, in other words, impute silliness to him. If interpretation is to make a lick of sense, silliness must be ruled out. The apparent need for maxims of interpretation is borne from the holistic nature of mental state ascription. Holism of the mental implies that many mental states get attributed at once, as a package deal. Absolutely unfettered interpretation would allow you to attribute whatever mental state you want, provided you compensate elsewhere.

This particular argument for a rationality requirement doesn't receive as much play in *Content and Consciousness* as does the one that I will call the Rylean argument, but it does play an increasingly prominent role in Dennett's writings as time goes on. For instance, he later writes,

> The assumption that something is an intentional system is the assumption that it is rational; that is, one gets nowhere with the assumption that entity x has beliefs p,q,r,… unless one also supposes that x believes what follows from p,q,r,…; otherwise, there is no way of ruling out the prediction that x will, in the face of its beliefs p,q,r,… do something utterly stupid, and, if we cannot *rule out* that prediction, we will have acquired no predictive power at all (Dennett 1978, 17, my italics).

According to Dennett, we need to "rule out" certain predictions. This is precisely why Quine, Davidson, and Lewis also hold fast to a principle of charity.[3] There are important differences between the three sons of Quine, of course. For one, they each have different opinions on the material one uses as input for the interpretive process. Davidson admitted publicly observable behavior, paying particular importance to the sentences that one asserts. Lewis allowed all physical facts, whether public or not, to be used as input for radical interpretation. Dennett can plausibly be read as allowing behavioral dispositions as well as the interpreted individual's (objective) goals or reasons as input.[4] Moreover, they all have different conceptions of what sort of norms of rationality are guaranteed. Still, they all agree that there is a need to

[3] Dennett also accepts Quine's argument in his (1989).

[4] Goals or reasons are characterized intentionally, which prevents Dennett from offering an account that fully naturalizes intentional descriptions to non-intentional descriptions. Note that taking reasons as input will not in itself guarantee rationality. Without a rationality constraint, it is still possible to interpret a person as irrationally ignoring what they have reason to do, or intending to do what they know is counterproductive to the attainment of their goals.

constrain interpretations with a rationality requirement in order to get any predictive or interpretive power whatsoever.[5]

In addition to Quine himself, a major source of historical support for this sort of rationality assumption came from formal decision theoretic models of economic behavior. Standard decision theoretic or game theoretic models, such as those of Von Neumann and Morgenstern (1944), are descriptive of human behavior only if humans act rationally and in accord with the dictates of the theory. Davidson, for one, was heavily influenced by Ramsey's "Truth and Probability" (1931). Ramsey gives a procedure for representing an agent's utilities and degree of beliefs in any proposition when simply given that agent's preferences; he then gives a representation theorem proving that if the agent's preferences satisfy certain requirements, the agent's degrees of belief will be coherent. Davidson took radical interpretation to involve something like Ramsey's procedure, and saw close affinities between Ramsey's procedure and Quinean radical translation. He writes, "Quine's solution resembles Ramsey's, in principle if not in detail."' (Davidson 1990, 319).[6] Dennett was less directly influenced by formal modeling (or at least, there is less textual evidence for its influence). He does at one point write that taking up the Intentional Stance involves interpreting an agent to have beliefs and desires "roughly as Bayes would have them" (Dennett 1978, 307), but formal decision theory has seemed not to have been a major influence. Still, it is worth noting that indeterminacy-reducing rationality constraints found wider appeal than simply among philosophers allied with Quine. Rationality requirements are what result from demanding that mental state ascription involve a procedure akin to Ramsey's. Choosing a formal theory that guarantees the ascription of rational beliefs is much the same as adopting a rationality constraint.

The sorts of considerations just mentioned make it *seem* like we need rationality constraints to get interpretation off the ground. But the arguments are not decisive. To my mind, the arguments fail to satisfyingly answer the following two questions:

1. Must some constraint on interpretation be a *rationality* constraint?
2. Is a constraint on interpretation really required *at all*?

[5] As an aside: it is worth nothing that although the philosophers above are interpretivists – they hold that the *content* of our mental states is determined through a process of interpretation – the apparent need for a rationality constraint hits non-interpretivists as well. Suppose that a computer or a brain contains an inscription written in Mentalese. Is this particular Mentalese sentence in a "belief box"? Or is it in an "imagination box" and the agent irrationally acts as if her imaginations are beliefs? Non-interpretivists find themselves facing the same problems that interpretivists do: they seem to require a rationality constraint to appropriately ascribe *attitudes* to an agent. We need to be careful and distinguish theories of semantic content from theories of mental attitudes with those contents. (This fact is sometimes glossed over by non-interpretivists. For instance, Fodor doesn't recognize this in his response to Stich's Mrs. T thought experiment, in which a woman assents to the claim that McKinley was assassinated while also being unable to say anything else related to assassination. Does she believe that McKinley was assassinated? Fodor should, I think, say she does not. She has the concept ASSASSINATED (fixed by asymmetric dependence), but it languishes in her head without playing a role in any of her beliefs. But this is not Fodor's response (see Fodor 1987, 62).)

[6] See Rawling (2003) for more on Quine and Ramsey's influence on radical interpretation.

I plan to argue that we already have principles that constrain indeterminacy, and an additional rationality requirement is neither motivated nor desirable. However, in order to talk about the principles that "we already have," I first need to unravel a persnickety issue that all-too-often complicates conversations about rationality constraints and interpretivism.

5.3 Types of Ascription

The rationality assumption is a constraint on theory construction. What sort of theory – and whose theory – requires constraint?

There are (at least) two sorts to consider. Firstly, individual human agents ascribe mental states to other agents. This is often called 'mindreading' or 'mentalizing'. One popular account of mindreading holds that we interpret other people around us by fitting our observations of them to a tacit folk psychological theory. The fitting of such a theory might involve an assumption of rationality. Secondly, philosophers and psychologists ascribe mental states to others by building, and subsequently applying, mature theories of the mind. This sort of theory-construction, too, might demand rationality constraints. Let's call these types of ascription *individual ascription* and *scientific ascription*, respectively. They are distinguished by who it is that does the ascription: the first is employed by individuals in real-world situations, and the second is employed by scientists and philosophers in the development of theories. Either investigation can have a descriptive or normative focus. One might be interested in how individuals actually do go about mindreading, or one can make suggestions about how people ought to mindread. Similarly, one can describe how psychologists actually do build theories that attribute mental states to observed actors, or one can offer suggestions about how their theories could be improved. Investigations into individual ascription are traditionally descriptive; investigations into scientific ascription are traditionally normative.

In *Content and Consciousness*, Dennett is clear that his concern is mental ascription of the second type. The goal is to build a mature theory of intentionality and mental states, and it is permissible to deviate from the terms of "ordinary" mental ascription. For instance, he writes, "the centralist makes his initial characterization Intentional, describing the events to be related in law-like ways using either ordinary, or semi-ordinary, or *even entirely artificial* Intentional expressions" (Dennett 1969, 41–2, italics mine).

The ground shifted somewhat when Dennett developed the Intentional Stance. The Intentional Stance became a piece of *individual* ascription: interpretation was now spoken as something that we *all* naturally do.[7] It is, of course, a legitimate

[7] E.g. "According to Intentional Systems Theory, [questions about the conditions under which a thing can be truly said to have a mind] can best be answered by analyzing the logical presuppositions and methods of our attribution practices, when we adopt the intentional stance toward something" (Dennett 2009, 339).

hypothesis that mindreading works through an application of a tacit theory of mind. However, building a psychological theory and mindreading are two separate enterprises, subject to different demands. Speed of processing is a worry in mindreading, for instance; the psychologist in her lab is under less time pressure.

The two enterprises became conflated in the literature. In "Mid-Term Examination: Compare and Contrast" (1989), Dennett takes a tour of his various philosophy of mind contemporaries, writing that "two chief rival" principles of interpretation have emerged: Normative Principles and Projective Principles. Normative Principles constrain interpretation by ascribing propositional attitudes that a creature *ought* to have; Projective Principles attribute the propositional attitudes that one supposes one *would* have in that very scenario. Dennett counts himself, Lewis, and Davidson among defenders of Normative Principles, and affirms that it all arose from Quine. Something strange has gone on here, however, for Projective Principles, with their egocentric focus ("interpret others as believing what *you* would believe in their shoes"), can only be understood as constraining individual ascription. To cast them as a competitor to the Normative Principles espoused by Quine, Lewis, and the Dennett of'69, suggests that these authors present their Normative Principles as also governing individual ascription, but this was not the case. Dennett, after all, suggests that a mature Intentional Systems Theory might invoke entirely artificial intentional expressions, formerly unknown to folk psychology. He can't be giving a theory about how we actually individually mentalize.

Dennett puzzles over the fact that Quine's *Word and Object* contains the seeds of both Normative Principles and Projective Principles. He resolves the potential conflict between the principles by arguing that for Quine, it did not matter much which principle yielded the actual propositional attitudes: since mental talk is a dramatic idiom that we employ simply for practical purposes, we can afford whatever indeterminacy is yielded by having two separate methods of ascription (344). I endorse a different solution: Quine presented the Projective Principle as part of a theory about how individuals actually understand the statements of others, and the principle of charity as a part of a theory about how linguists ideally ought to understand the statements of others. There is no conflict between the two principles because they are enlisted for two different projects. It is entirely consistent to be a simulationist with respect to individual ascription without being a simulationist with respect to scientific ascription: that is, while also being an interpretivist about the metaphysics of belief.[8]

Sometimes skeptics of rationality constraints admit that there is a need for something *like* a rationality constraint in order to act as an heuristic that can be used in real-time cognizing. This is not an admission that should be made if one is trying to determine whether we ought to invoke a rationality constraint when interpreting

[8] Goldman (2006) charges Dennett and Davidson with occasionally taking their theory of mindreading to be identical with their theory of the metaphysics of mental states, and their commitments to the metaphysics of mental states leads them to reject simulationism (a theory of *individual* mental state ascription) right off the bat.

others through a psychological theory. For instance, Cherniak offers a *minimal rationality constraint* because human beings are in "the *finitary predicament* of having fixed limits on their cognitive capacities and the time available to them" (Cherniak 1986, 8).[9] Bortolotti endorses an *intelligibility requirement*: "intentional behavior must be intelligible or amenable to rationalization" (Bortolotti 2009, 100), but she suggests that we should consider the interpreter's assumptions about intelligibility to be "flexible and revisable heuristics, not constraints. They are supposed to guide the interpreter and help her to ascribe intentional states with determinate content to a variety of subjects in a variety of situations" (107). The rest of their work makes it clear that they are really concerned with scientific ascription and the metaphysics of belief, so it is odd for them to discuss time-sensitivity and other concerns that clearly belong to the domain of individual ascription.

Now that we've established that the main project in *Content and Consciousness* is one of scientific and not individual ascription, an argument against the need for a rationality assumption can present itself.

5.4 Undoing the Quinean Lineage

We left our discussion of the Quinean lineage on a cliffhanger. Does Dennett have a good answer to the following two questions?

1. Must some constraint on interpretation be a *rationality* constraint?
2. Is a constraint on interpretation really required *at all*?

These are best dealt with in turn. Firstly, note that if the sole goal is to reduce indeterminacy of mental state ascription, it is far from obvious that a rationality constraint is the only constraint or assumption that would accomplish the task. It is one viable option, but there are others. One way to see this is to consider the argument from Ramsey's representation theorem. Ramsey showed that, given a preference ordering with certain features, humans can be formally represented as having rational and coherent degrees of belief, but this means nothing in itself, for they can also be formally represented as *irrational*. Zynda (2000) has shown that for any preference ordering that allows one to be representable as having degrees of belief that obey the laws of probability, that same preference ordering allows one to be representable as having of degrees of belief that *don't* conform to the laws of probability. In order to establish that humans are rational, it is not enough to simply establish that humans are representable as having consistent and rational beliefs; there are other representations that say otherwise.

This is just to say that the data are indeterminate without interpretation. But what's important is that the representations that lead to *ir*rationality are well-behaved, which means that the representation that guarantees rationality is only

[9] See also (Dennett 1987, 98).

one of many. What gives that particular interpretation a place of pride? It can't *simply* be its ability to reduce indeterminacy, because all sorts of representations have that feature.

What's even worse is that it appears that models that *don't* preserve rationality can actually be more predictive and empirically adequate. Since the original suggestion that unelaborated Ramseyan decision theory could be used as an empirical model of actual human decision-making (Edwards 1954), the claim has been steadily attacked; psychologists and behavioral economists have developed competing accounts of decision-making and competing research programs. Why should we think that the best formal theory of mental state ascription should be contained within the set of formal theories that guarantee rational beliefs? There are other formal models: some posit mental states other than belief and desire (such as intention or emotion); some do not assume that our preference ordering is transitive; some allow for unsharp probability functions. Perhaps a model that does not guarantee rationality will do a better explanatory job.

How does Dennett respond to apparent breaches of rationality in everyday life? After all, taking up the Intentional Stance involves interpreting an agent as having coherent and rational degrees of belief, but people obviously don't act exactly like perfect Bayesian agents all the time. Dennett accepts this, but he maintains that this doesn't imply the surprising fact that no one is a believer. He has two responses. Stich (1981) calls these "the hard line" and "the soft line" on rationality constraints.

On the hard line, the Intentional Stance is useful because people closely *approximate* rational agents. The property of *being a believer* is somewhat like the property *being a rabbit-shaped image* (Dennett 1991). Some images only vaguely resemble rabbits; others might be smudgy or pixellated. As the fidelity of the image goes down and noise is introduced, it becomes less of a perfect rabbit image, but it still has the same basic pattern that a perfect image would. People are, metaphorically, "smudgy images" of fully rational Bayesian agents. To ask whether a schizophrenic *really believes* that someone else has inserted thoughts into her head is akin to asking whether a shape in a smudgy picture *really is* rabbit-shaped. It's like a rabbit image in some respects but not in others – its status is indeterminate and there is no fact of the matter.[10] On the soft line, the form of rationality that is assumed by the rationality constraint demands less than perfect Bayesian consistency and coherence. For instance, it becomes rational to "satisfice" (to use Herb Simon's term). In Dennett's (1987) response to Stich, he adopts both strategies. So, upon seeing someone apparently act irrationally, we can either understand them by seeing them as approximating a rational being (and deviating slightly); or we can understand them as actually being rational according to some different standard.

The third strategy that Dennett does not adopt, of course, is just to give up on the assumption of rationality. Consider the hard line strategy: taking up the Intentional Stance just is representing or modeling an individual as having coherent degrees of

[10] Whether this account demands ontic vagueness is an open question; accounts of indeterminacy that are purely linguistic don't seem to capture what Dennett has in mind.

belief, and that people resemble perfect Bayesian creatures to some extent. Now, however, recall that there are other cognitive models waiting in the wings. Consider a new stance – a "schmIntentional Stance" – according to which individuals are represented or modeled by some different formal structure. Perhaps this representation assumes that we are predictably irrational whenever we reason about certain topics: perhaps it models us as systematically overestimating (or underestimating) the likelihood of events that would be bad (or good) for us. Perhaps it models us as having intransitive preferences (is this ruled out by the Intentional Stance?). Perhaps it posits various mental states that the Intentional Stance does not and to which it is difficult to apply folk notions of rationality. These models might very well do a better job of predicting human behavior.

Dennett often speaks as if, when an individual can't profitably be understood on the Intentional Stance, we need to plunge down to the design stance or physical stance. But why? Why not look for models of human psychology that are similar to (but distinct from) the one that you get by applying the Intentional Stance? We should not conclude that humans must be interpreted according to some psychological model just because they *can* be successfully interpreted according to that psychological model. There might be a more predictive model out there. We can update the Intentional Stance. That's what we do whenever cognitive psychology discovers new mental states.[11] The rationality constraint pushes us toward one of many possible interpretations of behavior. But in many cases, this means it pushes us away from interpretations that would be comprehensible and yield predictions.

Considering the second question (is a constraint on interpretation really required at all?) lets us go even further in questioning the need for the rationality constraint in theory building. Intentional Systems Theory models the mind, and we already have various maxims that regulate our theory construction. We do not need an additional constraint to reduce indeterminacy. Consider the observational data we acquire when building theories of physics. We take measurements, we construct atom chambers and run experiments, we build instruments, etc. The actual theory we construct is underdetermined by this data. We posit atoms and subatomic particles, but an evil demon manipulating all our observations will fit the data equally well. What prevents us from inviting in rampant indeterminacy in our commitments are certain epistemic principles or scientific virtues that guide our theorizing: simplicity, conservatism, scope, fecundity, and so on. If rationality were a constraint on mental state ascription, it would be serving as another such scientific virtue. It would be another such principle that we would use to reduce indeterminacy.[12]

[11] Two responses that Dennett might make here are responses that I will deal with in my discussion of the Rylean lineage in the next section. (A preview: they are that rationality is guaranteed by natural selection, so as evolved agents we are forced to make that assumption; and that the "schmintentional stance" is a discussion-changer: its declarations would be so remote from our ordinary mentalistic vocabulary that we could not properly call its posited states 'beliefs' and 'desires'.) All I am trying to establish here is that the need to reduce indeterminacy in ascribing mental states to our friend who is huddling under a tree in the thunderstorm does not *in itself* necessitate a *rationality* constraint, which is an argument that Dennett and others seem to make at times.

[12] The virtues listed above are some of those listed by Quine and Ullian (1970).

There is something very odd about this principle in that it is relative to a particular special science: psychology (and perhaps economics). No other sciences seem to require an additional virtue. This should make us suspicious of its necessity. In fact, it's not clear that the other virtues cannot do the job we want rationality to do. The problem is that ascribing irrationality to a subject is wholly uninformative. To interpret a person huddling under a tree as irrationally *not* intending to do so doesn't predict much else about them. Why will they say they are huddling under the tree? Would they huddle if they didn't want to huddle? The theory doesn't say. Attributing rationality and the intention to stay dry under the tree, on the other hand, offers up wealth of other information about their potential behaviors in various situations. This is very close to Quine's explanation: we attribute rational beliefs because we get predictive power by doing so. But the work here is not being done by an assumption that theories that postulate rationality are better: it's done by the assumption that theories with more predictive power are better.

Let's consider a version of Quine's field linguist, who, seeing speakers assent to an instance of ⌜p ka q⌝ when they assent to p and dissent from q, prefers to translate 'ka' as 'or' rather than as 'and'. Why should he prefer this hypothesis? On Quine's account, it would be because translating it as 'and' violates a requirement of rationality. Can we get the same result without appealing to such a constraint?

If we posit that 'ka' means 'or' and that the speaker is rational, we end up making all sorts of other predictions. For one, we anticipate that he will accept *any* instance of ⌜p ka q⌝ for any p or q. The hypothesis systematizes a whole lot of possible data about the speaker's dispositions. On the other hand, if we posit that 'ka' means 'and' and that the speaker is irrational, and if we don't have a theory about how the speaker is irrational, then we can't predict much else. We don't know how the speaker will respond to pretty much any instance of ⌜p ka q⌝. Thus, whatever scientific virtues push one to prefer simple and predictive systematizations of the facts will suggest a theory in which the agent is rational. We have a theory that tells us what can be expected when an agent is rational; claiming that an agent is irrational jettisons all those predictions. Consider: if a psychotic patient has the delusion that he is Napoleon, we can predict at least *some* things about his behavior (such as the fact that he will say that he is Napoleon). If we simply say that the agent has the irrational belief that he is Napoleon, then we should be hesitant to draw very few conclusions at all. We lose information. It's the epistemic virtues of predictiveness and systematization that keep us from attributing irrational beliefs, not a distinct rationality requirement.

Note that a rationality requirement can't be straightforwardly derived from predictiveness and systematization, because if we have a theory of how irrational actors will act, the most predictive, systematized, and empirically adequate theory might be one that interprets actors as irrational. Suppose we do come up with a theory of the speaker's irrationality. Suppose we notice that the speaker's behavior is altogether rationally consonant with 'ka' meaning 'and', but that the speaker tends to make errors when forming complex statements involving some particular sentence. We might then hypothesize that it's difficult for the speaker to reason about that sentence – maybe it introduces a lot of cognitive load. This hypothesis once again

lets us systematize the speaker's dispositions to assent: we expect that the speaker will assent to ⌜*p* ka *q*⌝ iff he assents to *p* and to *q* unless either *p* or *q* is one of the sentences identified to introduce cognitive load, in which case he dissents from the whole thing. The epistemic virtues should cause us to prefer a theory in which the agent is irrational if and only if the various irrational inferences the agent is disposed to make are patterned instead of piecemeal, and can be systematized into a theory of the agent's cognitive system that yields the patterns of irrationality.[13]

Sometimes it is argued that we should prefer models that assume humans to be rational because they are simpler than other models. Sober (1978) argues for this. Heil writes that it is useful to regard charity "as parsimony applied in the mental realm" (1994, 120). These sorts of warnings do not put any additional strictures on psychological theory-construction. We already have parsimony in the mental realm: it goes by the name 'parsimony'. Moreover, we don't want to *equate* charity with parsimony in the mental realm, because we cannot guarantee from the outset that the most parsimonious (or otherwise virtuous) theory will be the one with the result that people are rational. Thagard and Nisbett (1983) respond to Sober by presenting psychological evidence that people apparently behave irrationally in various domains; explaining away these apparent irrationalities will probably be less parsimonious than just positing a streamlined model that predicts irrationality in these domains. They present a moderate version of a principle of charity: "Do not judge people to be irrational unless you have an empirically justified account of what they are doing when they violate normative standards." This is not a bad general methodological principle (in psychology's current state). "Do not judge entities to be *X* unless you have an empirically justified account of how they can be *X*" is a reasonable scientific proscription whether building a theory of the mind or of tornados or of ducks. We have a simple theory of rational agents, we have some reason to think that rationality would be evolutionarily adaptive, and agents do seem to often be rational, so the rationality hypothesis is a reasonable default hypothesis. This is a far cry from saying that it is a constraint that cannot be overturned. If we find what appears to be systematic irrationality in people, then we needn't torture ourselves trying to interpret them as *really* being rational. We should just admit that the rationality hypothesis is no longer supported and then give it the boot.

I hope to have successfully challenged arguments that a rationality assumption is needed to do indeterminacy-reducing work because the work cannot be done by more standard scientific norms. If we interpret agents as rational because we are led to do so by scientific norms of predictiveness, systematization, and empirical adequacy, then rationality need not be a *constraint* on interpretation, nor need it play

[13] This account has affinities with Cherniak (1986), who argues that we don't only holistically ascribe mental states and language meanings: we holistically attribute mental states and the meanings of our words along with a theory of the agent's cognitive system. This is in order to account for the ascription of irrational inferences that are the product of memory constraints and computational difficulty or intractability. Cherniak, however, takes his project to be one of individual psychological ascription rather than the ascription of our best scientific theory, and still thinks that a constraint of minimal rationality is needed on top of all this.

any sort of role on the *input* side of psychological theory-building. It could be an *outcome*, or *finding*, of (current) psychology that agents are (largely) rational.

Consider, similarly, that it is an outcome of physics that there exist particles that have negative charge. We do not need to mandate anything like a negative-charge constraint on physics. It might be that a psychology that postulates rationality – or a physics that postulates electrons – makes better predictions, but this would only be contingently true, and not because of any necessary restrictions on theory-construction.[14]

This deals with the motivations for a rationality assumption that stem from Quine. The need to reduce indeterminacy in order to get psychology off the ground does not require anything that is unknown to the other sciences.

5.5 The Rylean Lineage

When Dennett entered his graduate studies at Oxford, ordinary language philosophy's Last Days of Empire were in full effect. When describing his time there, he emphasizes the atmosphere of disdain toward science that he experienced.[15] Attempted naturalizations of the mind were considered vulgar. Dennett broke from the tribe and auto-didactically immersed himself in psychology, neuroscience, and computer engineering, but even in so doing, he was moved by certain arguments of the anti-naturalists around him. The two books on intentionality that had the largest influence on him were Anscombe's *Intention* (1957) and Taylor's *The Explanation of Behaviour* (1964) (Dennett 1996). *Content and Consciousness* is studded with references to the two.

Dennett saw, in their anti-reductionist arguments, a recapitulation of Quine's arguments for the holistic nature and hence irreducibility of intentional discourse. While these arguments drove Quine to disparage mind-talk, in places advocating its dispensability and in other places treating it as pragmatic crutch that deserved scant respect, mind-talk was dead serious for the Oxbridgians. Their ordinary language

[14] There is a sense in which physicists do have something like a negative-charge constraint. If some feature of a theory has been pretty much conclusively established, scientists are free to dismiss theories that claim otherwise. Established physicists receive letters from all sorts of cranks who claim to have "disproved relativity," and these crackpots are rightfully ignored. The constraint in this case isn't a restriction on theory-building, but an heuristic used to guide the theorist's attention away from likely falsehoods. This does not always seem to be what Dennett has in mind when he speaks of a rationality assumption (for instance, when he argues that prediction could not get off the ground at all if it were not for an assumption of rationality).

Please note that in drawing the comparison between mental states and electrons, I do not mean to suggest that both are what Dennett calls 'illata' and that mental states are not personal-level states. Mental states are abstracta. Nonetheless, my comparison is apt because abstracta and illata are both potential objects of empirical investigation. Determining whether an agent has any particular personal-level state is an empirical matter. As I've been arguing, there's no compelling reason to think that empirical investigation into these sorts of states needs to involve a special sort of rationality assumption. (Note also that the positing of non-mental abstracta, such as centers of gravity, does not involve a rationality assumption.)

[15] Dennett (1996, 2012)

analyses of mentalistic terms proved attractive to Dennett. "The philosophy of mind initiated by Ryle and Wittgenstein is in large measure an analysis of the concepts we use at the personal level" (Dennett 1969, 95) and their sensitivity to the features of these concepts was crucial in the development of Dennett's theories. Ryle's notions of separate "logical categories" and the category mistakes that result from illicit admixtures of terminology from two different categories, foreshadows Dennett's construction of an Intentional Stance distinct from the Physical and Design Stance.

It's a conceptual analysis of mentalistic vocabulary that leads Dennett to his second version of the rationality requirement: it arises from the supposition that the meanings of mentalistic terms are fixed by their holistic connections with other mentalistic terms. Dennett points out the "conception causes pregnancy" is analytically true, because an event only counts as a conception if it causes pregnancy. Asking why a conception led to a pregnancy (rather than some other state) while using those terms is silly and unnecessary: the occurrence of the pregnancy is already entailed by there being a conception.[16] Dennett thinks mental vocabulary works in the same way. He writes,

> In Intentional explanation, on the other hand, the sequences of events are so characterized that the occurrence of a particular consequent action is explained by the occurrence of a particular antecedent, say a perception or a belief or intention, and there is no room for the question of why this consequent should follow this antecedent, and hence no room for any general law 'explaining' this sequence. For example, having said that my intention to leave was followed by my walking to the door, there is no room for the question: why should that result (as opposed to, say, opening my mouth or raising my arm) follow my intention to leave. The 'covering law' to the effect that all intentions to leave are followed by walking to the door is silly and unnecessary; the occurrence of my walking to the door has already been explained by citing my antecedent intention. In this way Intentional explanations assume the environmental appropriateness of the connections between antecedent and consequent (Dennett 1969, 37).

If you have a conception, then you certainly have a pregnancy, and this is guaranteed by the meanings of the terms. Similarly, if you have an intention to leave a room, then *ceteris paribus* and barring other mental states that would intervene, you'll move to leave the room; this is guaranteed by the meaning of the term "intention." If you acted irrationally instead of appropriately – if you opened your mouth or raised your arm – then you couldn't have had the intention in the first place. Whatever you had, it wasn't an intention to leave the room. To think otherwise would be to misuse the (ordinary language) word. For years, Dennett has presented various thought experiments to prompt the intuition that when rationality breaks down, we very much balk at ascribing beliefs to an agent: we don't know what to say. Let's draw another analogy with theories in physics. To be an electron, a subatomic particle must have certain features. It must have negative charge; it must have intrinsic angular momentum of 1/2, and so on. If some particle under observation does not display these properties, it isn't an electron. Similarly, for a

[16] This isn't actually true: 'in vitro conception' is in common use and not a contradiction in terms. (Admittedly, this is a cheap shot, as the technique was invented after the publication of Dennett's book. But this does go to show just how difficult it is to find analyticities.)

mental state to play a belief-role, it might need to stand in rational relations with other mental states.

The claim that beliefs are constitutively rational can be read in two ways, and they are not always distinguished. Firstly, one might mean that the *process of interpretation* involved in mental state ascription is constrained by a principle that guarantees the rationality of the interpreted agent. Alternately, one might mean that it is characteristic of the *functional role* of belief that it is rational: if a mental state doesn't play the role of a rationally formed and maintained belief that motivates behavior in a rational way, then it doesn't play the role of a belief. One way to think of this is that on the first thesis, rationality is a condition on the interpretive process. On the second, rationality is a mandated feature of the outputs of the process of interpretation. The first sort of rationality constraint is Quinean, and the second is Rylean.

5.6 Undoing the Rylean Lineage

Suppose we grant that the meaning of 'intention' in everyday folk language does, in fact, imply that individuals act appropriately on their intentions. Why must Intentional Systems Theory hang onto the meanings given to us by folk theory unaltered?

I am not driving toward eliminativism; I'm not suggesting that we replace belief-desire psychology with something radically different. My goal is less contentious. I am simply pointing out that once we separate the project of explaining individual ascription from the project of scientific ascription, we should recognize that it is perfectly admissible to make modifications to folk theory if it gains us predictive and explanatory power. Dennett himself does this: recall his claim that a successful Intentional Systems Theory might describe mental events using "ordinary, or semi-ordinary, or even entirely artificial Intentional expressions" (42). In a chapter of *Brainstorms*, he introduces *opinion* as a novel sort of propositional attitude, and touts it as "a *reform* of our ordinary concept of belief" (Dennett 1978, xxii). It's true that opinions were introduced in order to *preserve* rationality: when an agent says P, and it would render him irrational were he to believe P, we can say instead that he merely has the opinion that P. But the damage is done: folk psychology is up for amendment if in the service of constructing a better theory. Why not think that the features of folk explanation that presume appropriateness are similarly up for grabs? The simple fact that folk psychological terms assume rational relations does not in itself say anything about whether the terms of a mature theory ought to similarly assume rational relations. We might find it best, at some point, to adopt the schmIntentional Stance instead.

Thus, even if the terms of folk psychology analytically ensure the rationality of any agent they are attributed to, this would not, in itself, restrict future theory-building. We regiment folk terms all the time in all the sciences; why are these terms sacrosanct? One might think that it is just central to the meaning of 'intention' that

it implies rational relations to other mental states. If this were true, then amending intention to be intention-like really would be considered a version of eliminativism. To my ears, this sounds like a semantic dispute over what states merit the name 'intention'.[17] Arguing over whether an irrational intention is an intention does not sound much different to me than arguing whether a wrap is a sandwich.

5.7 Preserving Intentional Systems Theory

I believe that electrons exist; I also believe that they have negative charge. I do not think that we are in much danger of a future generation discovering that there are no electrons. Imagine, then, that I encountered someone who believed in a *negative charge requirement* on the construction of physical theories. He presents me with the following two arguments: firstly, it is an indeterminacy-reducing constraint on theory-building that physics posit a subatomic particle with negative charge; secondly, it is a central part of our current concept that electrons have negative charge, and the concept is too useful and predictive to ever want to give up.

This "negative charge requirement" is wholly unneeded. The existence of subatomic particles with negative charge was a discovery, not an a priori condition on scientific inquiry. If anything, having this sort of requirement stunts potential scientific investigation: on the remote chance that a there is a more virtuous theory waiting in the wings that would dispense with negatively charged particles, the requirement would have us dismiss it out of hand.

Rationality requirements are in much the same boat. I think we have good reason to suppose that our mental states are (mostly) appropriate and rational, but this is a well-established *discovery*, not a condition on all future psychologizing. One argument Dennett makes for rationality requirements which I haven't mentioned until now appeals to natural selection. Having mostly true and rationally-formed beliefs is conducive to fitness, so we should expect our attitudes to be rational. I think this is a good argument.[18] However, I have a hard time seeing how it could act as an argument for a rationality constraint on mental ascription. Our evolved nature is a source of evidence that should cause us to *expect* our attitudes to be rational, but this evidence could plainly be defeated by other sources of evidence. Perhaps we will find that it was fitness-conducive in our primitive niche for us to be overly credulous or skeptical, or to be subservient to authority – biases that are harmful in our current environment. The various cognitive biases that psychologists discover do provide *some* evidence that we are irrational; they can't always be written off (as performance errors or whatnot) in allegiance to an unshatterable rationality assumption.

Does it drastically damage Intentional Systems Theory if we scrap the rationality requirement and simply replace it with a claim that we have a lot of *good evidence* that our mental states are rationally arranged? I can't see that it does. Dennett still

[17] See Stich (1996) for more on these tricky semantic issues.

[18] Pace Stich (1985).

has the ability to claim that mentalistic vocabulary is holistic and irreducible. He can still hold that we are goal-oriented, sensitive to reasons, and have his version of free will worth wanting. He can also still hold that *individual* ascription involves the use of a rationality assumption: it's just one of many heuristics that we might use in order to enable real-time mentalizing.

On the other hand, it might be thought that I haven't made much of a change. Is progress really made by saying that, instead of there being a *rationality assumption* on ascription, it is epistemically *safe to assume* that minds are mostly rational? Yes, I think so. It's a small point, but an important one: by removing rationality as a condition on all theories of mind, we remove a barrier that could influence or stand in the way of creative theory construction. Philosophy of mind has of late been replete with proposals for new attitude types much like Dennett's own opinions, from Gendler's aliefs (2008) to Egan's bimagination (2009) to Schwitzgebel's in-between beliefs (2001) to Frankish's superbeliefs (2004). These are exciting and creative times. I worry that the rationality constraint is too aprioristic, and it will dissuade us from imaginative reform of our cognitive theories.

References

Anscombe, E. (1957). *Intention*. Cambridge, MA: Harvard University Press.
Bortolotti, L. (2009). *Delusions and other irrational beliefs*. Oxford: Oxford University Press.
Cherniak, C. (1986). *Minimal rationality*. Cambridge, MA: MIT Press.
Davidson, D. (1982). Psychology as philosophy. In *Essays on actions and events* (pp. 229–238). Oxford: Oxford University Press.
Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy, 87*(6), 279–328.
Dennett, D. (1969). *Content and consciousness*. London: Routledge and Kegan Paul.
Dennett, D. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge: Bradford Books.
Dennett, D. (1987). Making sense of ourselves. In *The Intentional Stance*. Cambridge, MA: MIT Press.
Dennett, D. (1989). Mid-term examination: Compare and contrast. In *The Intentional Stance*. Cambridge, MA: MIT Press.
Dennett, D. (1991). Real patterns. *Journal of Philosophy, 88*(1), 27–51.
Dennett, D. (1996). An overview of my work. In K. Ouyang & S. Fuller (Eds.), *Contemporary British and American philosophy*. New York: Nova Scientific Publishers.
Dennett, D. (2009). Intentional systems theory. In B. P. McLaughlin (Ed.), *The Oxford handbook of philosophy of mind*, chapter 19 (pp. 339–350). Oxford: Oxford University Press.
Dennett, D. (2012). Daniel Dennett: Autobiography (part 1). *Philosophy now*.
Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51*(4), 380.
Egan, A. (2009). Imagination, delusion, and self-deception. In T. Bayne & J. Fernandez (Eds.), *Delusion and self-deception: Affective and motivational influences on belief formation* (pp. 263–280). New York: Psychology Press.
Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.
Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.
Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy, 105*(10), 634–663.

Goldman, A. (2006). *Simulating minds*. Oxford: Oxford University Press.
Heil, J. (1994). Going to pieces. In G. Graham, & G. L. Stephens (Eds.), *Philosophical psychopathology* (pp. 111–134). Cambridge, MA: MIT Press.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, MA: Cambridge University Press.
Lewis, D. (1974). Radical interpretation. *Synthese*, 27(July–August), 331–344.
Quine, W. V. (1960a). Carnap and logical truth. *Synthese, 12*(4), 350–374.
Quine, W. V. (1960b). *Word and object*. Cambridge, MA: MIT Press.
Quine, W. V., & Ullian, J. S. (1970). *The web of belief*. New York: Random House.
Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Routledge and Kegan Paul.
Rawling, P. (2003). Radical interpretation. In K. Ludwig (Ed.), *Donald Davidson, contemporary philosophers in focus* (pp. 85–112). New York: Cambridge University Press.
Schwitzgebel, E. (2001). In-between believing. *Philosophical Quarterly, 51*, 76–82.
Sober, E. (1978). Psychologism. *Journal for the Theory of Social Behavior, 8*, 165–191.
Stich, S. (1981). Dennett on intentional systems. *Philosophical Topics, 12*(1), 39–69.
Stich, S. P. (1985). Could man be an irrational animal? *Synthese, 64*(1), 115–135.
Stich, S. P. (1996). Deconstructing the mind. In *Deconstructing the Mind* (pp. 3–90). Oxford: Oxford University Press.
Taylor, C. (1964). *The explanation of behaviour*. London: Routledge and Kegan Paul.
Thagard, P., & Nisbett, R. E. (1983). Rationality and charity. *Philosophy of Science, 50*(2), 250–267.
von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science, 67*, 45–69.

Chapter 6
Dennett's Personal/Subpersonal Distinction in the Light of Cognitive Neuropsychiatry

Sam Wilkinson

Abstract In this paper, I examine Dennett's personal/subpersonal distinction. However, there are two versions of the distinction within Dennett's work: an earlier and a later one. My aim is to clarify both versions of the distinction and examine them in the light of a particular enterprise, namely cognitive neuropsychiatry. In particular, the two versions of the distinction cast delusional subjects in very different lights. According to the later distinction, delusional subjects fail to have mental states attributable to them to the extent that they cannot be predicted by the intentional stance. According to the early distinction, personal-level explanations can, in principle, be used to account for this unpredictability.

In this paper, I examine Dennett's personal/subpersonal distinction. However, there are two versions of the distinction within Dennett's work. The earlier one, which is (unsurprisingly) closer to what we find in Ryle's work, is to be found in *Content and Consciousness* (1969).[1] The later version, which first appeared in Dennett (1971) and was subsequently published in *Brainstorms* (1978), comes hand-in-hand with his intentional stance. My aim is to clarify both versions of the distinction and examine them in the light of a particular enterprise, namely cognitive neuropsychiatry.

I proceed as follows. I present Dennett's distinction, first in a general way that applies to both the early and late versions of the distinction, and then I highlight the difference between the two versions. Then I look at cognitive neuropsychiatry, in particular in the case of delusions (obviously false or unwarranted beliefs) occurring in the context of brain damage, and the use it makes (or ought to make) of personal and subpersonal kinds of explanation. I end by reflecting on what cognitive neuropsychiatry reveals about the two versions of the distinction. In particular, the two versions of the distinction cast delusional subjects in very different lights. According to the later distinction, delusional subjects fail to have mental states attributable to them to the extent that they cannot be predicted by

[1] I say "unsurprisingly" because *Content and Consciousness* was based on Dennett's D Phil thesis, which was written under Ryle's supervision.

S. Wilkinson (✉)
Durham University, Durham, UK
e-mail: slj_wilkinson@hotmail.com

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_6

the intentional stance. According to the early distinction, personal-level explanations can, in principle, be used to account for this unpredictability.

6.1 Introducing Dennett's Distinction

It is crucial to note the nature of Dennett's distinction, in both its guises. It is not, primarily, a distinction between concrete phenomena. If one is to be true to the distinction as Dennett introduces it, one strictly speaking ought not (as many philosophers and psychologists unfortunately do) speak of personal or subpersonal *processes* or *states*.[2] The predicates "personal" and "subpersonal" apply to vocabularies, to levels of discourse, and, by extension, to explanations. The purpose of the personal/subpersonal distinction is to guard those speaking of mental phenomena from making category mistakes. In other words, one should not mix up personal and subpersonal vocabulary, especially when constructing explanations. It is a mistake (not in the sense of being straightforwardly false, but rather it is confused) to identify a pain or a belief with a particular neural state or process. It is *persons* who are in pain, *persons* who believe. This much applies to both the early and the later versions of the distinction. However, we will see later how these two versions differ.

Although one can hold some version of the personal/subpersonal distinction without this, I find it enlightening to think of the early Dennett as being heavily influenced by Ryle.[3] This influence manifests itself as a focus on language use. If language users are to use a language successfully they must use language correctly. Incorrect use of language, at least in a context where you are trying to make assertions (for, clearly, there are non-assertive uses of language) does not mean that you risk saying something false, but rather that you will be saying something that is meaningless, capable neither of truth or falsity (or rather not "saying" anything at all, in a different sense of "saying"). The emphasis is on the use we actually make of language when we use it correctly.

Dennett makes his adherence to this approach clear, early on in *Content and Consciousness*: "We have the mental language, and *since the suggestion that all the things we say in the mental language might be false is incoherent*, we also have the truths expressed in mental language" (1969, p. 19, emphasis added). Unlike many philosophers, notably eliminativists, the fact that we (most of us, most of the time) correctly use mental language, and say things that are true with it, is not something to be questioned. Rather, it is a starting point, a *datum*. The task is not to question this, but rather to clarify how this happens. To use Dennett's own example, the task is to elucidate the conditions under which we can (and it is assumed that we can)

[2] Many, for example, equate the personal/subpersonal distinction with the conscious/unconscious distinction, which is disastrous.

[3] Many philosophers (e.g. Fodor, Davidson etc.) will accept that different sciences must employ their own vocabulary, and therefore may well accept some version of distinction that is language-based, without this Rylean focus on language use.

truthfully say "Tom is thinking of Spain" rather than finding a *thing* that is Tom's thought about Spain (that is first referred to, then predicated of). This is what Dennett means when he speaks of mental language being "non-referential" (1969, pp. 89–90).[4] As a result, "we absolve the scientist from the responsibility of discovering physical events, states or processes which deserve to be called thoughts, ideas, mental images and so forth."

This underlying emphasis on language use, (i) allows that we can say things that are true without there being a spatiotemporal thing to which we are referring (Dennett illustrates this in Chap. 1 with his example entity of "a voice", when we say that someone has a strong voice), and (ii) places particular kinds of constraints on what counts as meaningful use of language. These constraints, which Ryle applies in *The Concept of Mind* (1949) to the Mind-Body problem, Dennett applies in *Content and Consciousness* to the scientific study of the mind. The question "Here is the brain, but where is the mind?" – as meaningless to Ryle as the question, asked of a university tour guide, "You've shown us the university buildings, but where is the university?" – is analogous to the question "I see what goes on neurally when someone is in pain, but where's the pain?" This is mixing up personal and subpersonal kinds of discourse. It is to prevent these kinds of questions from being asked, or thought meaningful and answerable, that the personal/subpersonal distinction is introduced.

With this in mind, let us look more precisely at how Dennett characterizes his distinction in *Content and Consciousness*.

6.2 Dennett's Early Distinction

Dennett introduces his distinction (1969, p. 90) with regard to pain. If we render someone's behavior intelligible by saying that she is in pain, there is a sense in which "we have said all there is to say within the scope of this vocabulary" (1969, p. 93). We can look for "alternative modes of explanation" and "turn to the *subpersonal* level of brains and events in the nervous system. But when we abandon the personal level in a very real sense we abandon the subject matter of pains as well" (p. 93–94). More to the point, the early Dennett thought, we change the kind of discourse that we are engaged in. Furthermore "[t]he only sort of explanation in which 'pain' belongs is *non-mechanistic*" (ibid., emphasis added).

So there are two distinct claims Dennett makes here. First, there is the claim that there are two kinds of discourse, and second, that one of these kinds of discourse is used in providing non-mechanistic explanations. It is important to see how we get

[4] In the light of this, one can see both realists and eliminativists as being two sides of the same "referentialist" coin. Both think of mental language as referential. Fodor thinks that it is referential and true, and therefore there must be things (e.g. beliefs) that we refer to, whereas eliminativists (e.g. Churchland 1986) take it to be referential but as failing to achieve reference, since states like beliefs do not literally exist.

from one to the other, for it seems possible that two different and incommensurable kinds of discourse could be used in making the same kinds of explanations, e.g. mechanistic explanations. In order to see how Dennett gets from one claim to the other, we need to examine a now widely acknowledged point concerning the metaphysics of mind, and a (less widely acknowledged) point about the way personal-level explanation works.

The former point is as follows. The objective conditions under which I can truthfully say "Tom is in pain" may coincide with the conditions under which a certain, say, neurological statement is true. Although the conditions in which those statements are true may coincide in this world, functionalism has taught us that they need not coincide in other possible worlds, for example, where Tom (or the organism in pain) is physically constituted rather differently. In short, pain is "multiply realizable". This functionalist lesson is widely accepted. The latter point about explanation is perhaps less so, and it is as follows.

Sticking to *this* world, so to speak, the claim is that we also wouldn't want to replace the word "pain" with a neurological description in an explanation. In particular, neurological statements and pain statements do not have the same explanatory power. I will illustrate this point first, and flesh it out later. In answer to the usual (interpretation of the) question "Why is Tom wincing?" if you were to describe, however fully and accurately, his neurological state, that would not be a good answer to *that* question. Not only would I not be able to make sense of it (because I am not sufficiently knowledgeable about neurology), but also, even if I could, it would not be an answer to the question I had asked. The answer, "Because he is in pain", on the other hand, is, however simple, precisely the answer I'm looking for. I am looking for an answer – an explanation – that renders a person's behavior intelligible, not one that renders the motion of a causal system tractable.

What is this intelligibility? Suppose we think of explanation as providing a certain kind of answer, a genuinely satisfying answer, to a particular kind of question. What determines the kind of question is the nature of the explanatory concerns of the person asking the question. Sometimes these concerns are mechanistic and we want to know *how* a causal system functions. But sometimes we want to know *why* someone has behaved or acted in a certain way. In this case, we are looking for a personal-level explanation that renders the subject intelligible. In such a situation we come to understand the subject, and a certain explanatory concern is fully satisfied. Their behavior ceases to perplex us.

Indeed in holding this, Dennett explicitly pays tribute to Ryle and Wittgenstein: "[…] the lesson to be learned from Ryle's attacks on 'para-mechanical hypotheses' and Wittgenstein's often startling insistence that explanations come to an end rather earlier than we had thought is that the personal and sub-personal levels must not be confused." (p. 95)

The fact that explanations provided by personal-level discourse are non-mechanistic is more clearly seen later in the book, when Dennett moves away from pain and looks at propositional attitudes and rational agency. With beliefs and actions, we often ask intelligibility-demanding questions of one another: "Why do you believe that?" "Why did you do that?" What is somewhat clearer than in the

case of pain is that this is asking a very particular kind of question, and one that requires a very particular kind of answer. This answer is commonly called a *rational* explanation and it, like the explanation that appeals to pain, is non-mechanistic. Now, this is not, in itself, to deny that reasons are causes. This is especially the case if one subscribes to a counterfactual theory of causation. Insofar as it is true that if I hadn't believed or desired what I did, I wouldn't have done what I did, reasons clearly are causes.[5] The point is that they do not explain in virtue of elucidating a mechanism. It is not in virtue of elucidating a mechanism that my explanation satisfies the person who has asked it of me. Nor is it the case that my talk of pain or belief is a mere place-holder (or "mechanism sketch", as Piccinini and Craver (2011) put it) that I will want to fill once I know the right mechanistic details. My personal-level answer to the personal-level question is already complete, and the relevant mechanistic details would not provide an answer to that question.[6]

The following illustration will seem familiar given what I said about pain, but the point it illustrates is more clearly seen. If you ask me, "Why did you raise your hand?" and I answer, "Because I wanted to ask a question" that's normally a satisfying explanation. If I tell you a full physiological story about what happened up until the moment my hand went up, that may be interesting, but it's not an answer to *that* question. A description of a causally related sequence of events is not what you asked of me. You were after a *justification*, a *reason*. And this is not mechanistic. The same applies when you ask the question "Why do you believe this?" You are after *reasons* for my belief, not any mechanistic story.

If pain seems less clearly illustrative of the distinction, one may wonder why Dennett introduces the distinction with reference to it. However, it is illustrative to reflect on why he does so. One reason may be that it is vivid (and perhaps this has lead some to erroneously equate "the personal-level" with "the subjective" or "the phenomenally conscious"). More importantly, I would suggest, it is because the "physiology of pain is relatively well understood" (p. 90). On the other hand, the neural underpinnings of *belief* (if it even makes sense to speak of such a thing, which I doubt) are not so well understood. He wants to show that our empirical knowledge of neural phenomena, though extremely interesting and important, doesn't affect our personal-level discourse: no matter *how* well we know the physiology of pain, we will never be really talking about *pain* in the personal-level sense. It is not the physiology of pain, or the neural underpinnings of my belief, that provide *grounds* for wincing or belief.

Another useful thing to note about Dennett's use of pain as his example phenomenon, is that personal-level explanation cannot be the same as rational explanation, at least not in one sense of "rational". "Rational" has a categorical and an evaluative sense. In the categorical sense, the opposite of "rational" is "non-rational". You can talk about rational or non-rational entities, and processes. Thus, a rock, or a chair, is a non-rational entity, whereas a person is a rational entity. An action is a rational

[5] Note that pain is a cause in a similar sense: If I hadn't had the pain, I would not have winced.

[6] For a more in-depth treatment of different kinds of explanation in terms of different kinds of question, see Wilkinson 2014.

process, a blink reflex is not. In the evaluative sense, the opposite of "rational" is "irrational" and these evaluations only make sense when applied to things that are categorically rational. Thus you don't have irrational rocks, but you have irrational persons, you don't have irrational blink reflexes, but you do have irrational actions. Now think about pain, and the behavior that it can give rise to. The presence of the pain itself is non-rational. However, insofar as a pain can interfere with the achieving of goals, it can yield behavior that is, to some extent, irrational.[7] If somebody pinches me while I'm at the Opera and I shout "Ouch!", thereby disrupting the performance and embarrassing myself, I have behaved in a way that goes against my goals and intentions. In such a case we are not in a position to give a rational explanation of action. You cannot explain my shouting "Ouch!" in terms of my beliefs, desires and intentions, since it precisely goes against my beliefs, desires and intentions. And yet we can give a personal-level explanation of the behavior. My behavior (e.g. shouting "Ouch!") is rendered intelligible to us when we are told that I was pinched in a way that badly hurt. Ascribing pain renders pain behavior intelligible. In one sense of the why-question: "Why did you say "Ouch"?", the answer "Because that hurt!" is a complete and satisfying answer (in a way that a physiological answer would not be). Similarly for belief, if you ask, "Why do you believe that?" I ought to say, "Because James told me that, and I trusted him." My belief is intelligible to you as a result of this story. The explanation of the belief is a rational story as well as a personal-level story. The explanation of the "Ouch!" is only the latter. Neither is a mechanistic story. It is not thanks to understanding any mechanism that you come to find my shout or my belief intelligible.

Notice that if we can render the pain behavior intelligible by appeal to pain, then we can at least imagine situations where that behavior is not intelligible, namely by removing that which grants intelligibility, in this case pain. Of course, other things can grant intelligibility to indistinguishably similar overt behavior. Someone might wince without being in pain if they are acting on stage, but there we can say that they *want* to communicate that their character is in pain and *believe* that this is the way to go about it. But suppose that somebody winces for *no reason at all*; they are not in pain, not acting, not just being silly etc. it's just pure behavior (we may say that it's a tick). This, indeed by stipulation, is not amenable to personal-level explanation. In such a situation our explanatory concerns may shift quite radically, and lead to an inquiry aimed at giving us a subpersonal, physiological story (by what mechanism is the tick caused?).

Let me sum up this section. The early personal/sub-personal distinction involves a combination of two claims. The first claim is that personal-level vocabulary is irreducible. The second claim is that this vocabulary is used to give non-mechanistic explanations, and, more specifically, to give explanations that render the subject intelligible. It may be illustrative to consider this as corresponding to causal/mechanistic and justificatory uses of the word "because". A causal/mechanistic use renders causally unmysterious a particular *explanandum* (e.g. a state or event):

[7] Of course, pain can also be implicated in perfectly rational action. I believe that I am in pain, I don't want to be in pain, and I believe that doing this will alleviate my pain.

"There is a hole in the Ozone layer *because* of free radicals in the atmosphere". A justificatory use provides grounds for action ("I bought oregano *because* I wanted to make Provencale Chicken") and belief ("I believed that Tom was home because I saw his car in the driveway"). As we saw, it can also be used in explaining pain behavior. "I said "ouch" because that hurt!" (This does not seem to be a causal use of "because").

We will see that Dennett's later version of the distinction, in a sense, drops a commitment to the second claim.

6.3 The Later Version and the Intentional Stance

A few years later, Dennett maintains the distinction, and it remains indispensable, but in a somewhat different way. One way of seeing the difference with the early Dennett is that the concern is now less with *explanation* (especially in terms of intelligibility), but more with *prediction*. Adopting the intentional stance (i.e. using the personal-level vocabulary of beliefs, desires etc.) is a "strategy for predicting the future behavior of a person" (1981, p. 557), a strategy contrasted, in method but not in aim, with other strategies and stances, such as the "physical stance" and "design stance". So it is, first, a useful shortcut, and an indispensable one for creatures like ourselves. We have neither the knowledge nor the time to predict each other's behavior using the physical or design stances. But it is more than merely indispensable in this weak, contingent, sense of being indispensable to us because of our cognitive limitations. Dennett makes this clear in the following example (attributed to Robert Nozick). Suppose we remove these limitations. Suppose "some beings of vastly superior intelligence – from Mars, let us say – were to descend upon us […] suppose, that is, that they did not need the intentional stance – or even the design stance – to predict our behavior in all its detail." (Dennett, 1981, p. 562).

The question then is: do these Martians miss out on anything in failing to use the intentional stance, the personal-level vocabulary of beliefs, desires etc.? According to Dennett, although they "might be able to predict the future of the human race […] if they did not see us as intentional systems, they would be missing something perfectly objective: the *patterns* in human behavior that are describable from the intentional stance, and only from that stance, and that support generalizations and predictions." This point about the irreducibility of the intentional stance, these patterns, is a point about the fineness of grain in the explanation. Something is missed by the Martians because they are not operating at the relevant coarseness of grain. They may, as Dennett points out, be able to predict the exact motions of the fingers and the vibrations of vocal cords during an instance of a stockbroker buying shares in General Motors, but if they fail to see

> that indefinitely many *different* patterns of finger motions and vocal cord vibrations – even the motions of indefinitely many different individuals – could have been substituted for the actual particulars without perturbing the market, then they would have failed to see a real pattern in the world they are observing (ibid.).

"Seeing patterns" in any system is about understanding what is a significant kind and level of variation and what is not.

Indispensable as this is, it is (like the physical and design stances) still concerned with behavioral prediction rather than with explanation in terms of intelligibility. The fineness of grain, the fact that our super Martians miss out on an understanding of the relevant patterns, and that these patterns are real; all of this is absolutely correct, and important. However, it abandons something central to the early version of the distinction. We don't simply use personal-level vocabulary to predict behavior (at the relevant fineness of grain); we use it to justify our actions and beliefs, and to render the actions, beliefs and behaviors of others intelligible. And that is why they are used in a special kind of non-mechanistic explanation. When this fails and others become *unintelligible*, we aren't simply bemoaning the fact that this de-rails our predictive power over a causal system. Indeed that predictive power could be acquired by "turning to the subpersonal level of […] events in the nervous system". What we are bemoaning is that the people in question have become perplexing (and in such a way that a causal understanding alone would make them no less perplexing).

One major consequence of the abandonment of this focus on intelligibility in favor of predictability is that the intentional stance (and hence the version of the personal-level that is tied to it) can be applied, not only to non-humans, but also to non-biological beings. There is no line to be drawn. Indeed Dennett is famous for claiming that the intentional stance is just as metaphorical for humans as it is for chess computers. The earlier Dennett, Ryle's disciple, would certainly have agreed that there is no *metaphysical* line to be drawn. Certainly, human beings are just biological machines. However, he would have granted that there is a crucial conceptual line to be drawn. Given our use of personal-level vocabulary, and the way in which it acquires its meaning and enables us to say things that are true, there is nothing metaphorical in saying that a person believes something (although there is something metaphorical in saying that a chess computer believes something).[8] Nor is there anything metaphorical about saying that someone is in pain and that that is why they have behaved in the way that they have (it is, for example, an undeniable fact that I am not in excruciating pain right now, and that is reflected in my behavior). It is central to the early Dennett that this can be as literally true as anything can be. There is also nothing metaphorical about saying that someone did something for these reasons, and thereby acted rationally (or irrationally).[9] This, too, can also be literally true. Furthermore it forms the basis, not only of many of our daily interactions with other (predominantly healthy) human beings, but also of some of our categorizations of subjects with mental disorders.

With this in mind, let's look at cognitive neuropsychiatry, and see what it may suggest about these two versions of Dennett's distinction.

[8] Whether it is metaphorical to say that animals (and which ones) believe is less obvious.

[9] Or *think* that they did something for certain reasons, but actually did not (see, e.g. Gazzaniga 1995 on confabulation in split brain patients).

6.4 The Personal/Subpersonal Distinction and Cognitive Neuropsychiatry

Cognitive neuropsychiatry is, roughly speaking, the attempt to understand mental illnesses in terms of our best models of normal cognitive functioning.[10] Why we would want to undertake such an enterprise is obvious enough. Since mental health depends, at least in part, on the brain's functioning properly, our understanding of the brain and cognition can contribute to our understanding of mental illnesses. Indeed one might even say that as our knowledge of how the brain works advances through the fields of neuroscience, cognitive psychology and neuropsychology, it is our duty to integrate this knowledge into our understanding of mental illness as best we can, with the ultimate aim, of course, that we will be better placed to treat the mentally ill. But how is this knowledge to be integrated? What are the constraints on such an integration? These questions concerning the nature, extent and possible restrictions on such an integration are arguably concrete applications of the very same issues that are central to the personal/subpersonal distinction. Indeed the subject-matter of neuropsychiatry constitutes particularly useful phenomena for testing versions of the personal/subpersonal distinction and their applications, since they seem to instantiate more or less direct interaction between paradigmatically subpersonal phenomena (e.g. brain damage, dopamine dysregulation) that can be described in various different ways by different disciplines (neuro-anatomy, neurobiology, cognitive neuropsychology, computational neuroscience etc.), and paradigmatically personal phenomena (e.g. certain beliefs, experiences, emotions, actions). How do the mechanistic, subpersonal, explanations provided by the sciences of the brain relate to the beliefs, desires, intentions, experiences, emotions, etc. of persons (especially those with mental disorders)?

Now here is an eminently plausible answer. The brain sciences can give us a mechanistic explanation, can tell us *how come* the subject, qua causal system, comes (for example) to be experiencing what she is experiencing. Then, once the nature of the experience is adequately characterized, we may then be in a position to understand the subject better, namely, to render their beliefs, desires, actions (or even mere behavior) intelligible. Something like this is already implicit in much of the best neuropsychiatry. Let us look at some examples.

Take a primary symptom of schizophrenia, the phenomenon of delusions of control. Someone with delusions of control claims that her actions are being controlled by an external force or agent.[11] There are two very different kinds of question we can ask about this. For example, you could ask either:

---

[10] There is clearly two-way interaction here. Mental disorders can give us insights into how cognition functions generally.

[11] It is illustrative, as Frith et al. (2000) do, to contrast delusions of control with anarchic hand syndrome. In the latter, subjects end up not behaving as they would like to (e.g. they want to get dressed, but the anarchic hand unbuttons the shirt as the subject tries to button it up) but the actions do feel self-produced. In the former, subjects act in accordance with their intentions, but feel like their actions aren't self-produced. For example, they see that their hair needs combing, and comb their hair, but the action of combing their hair does not feel self-produced.

(A) "How come this causal system (this human animal) behaves the way it does?"

Or:

(B) "Why does the subject deny that she is causally responsible for her actions?" (Or, the more general version of this: "What could make someone deny control of their own actions?")

Note that with a sufficiently complete answer to A, we could predict the subject's behavior *qua causal system*. But this need not answer B. It might, and should, *help* us answer B, but if A expressed our only concern, addressing that concern wouldn't *ipso facto* address the concern expressed in B. It would take a further bit of work to address B. And one could also answer B satisfactorily without even coming close to being in a position to answer question A (we give, and have given for centuries, B-type explanations without *any* understanding of the mechanisms that A-type questions allude to).

One nice example of a hypothesis that tries to answer both A and B-type questions is Chris Frith's account of the primary symptoms of schizophrenia (see, e.g. Frith 1992, Frith et al. 2000), which views them as the upshot of deficits in self-monitoring.

Perhaps the first theorist to make use of the notion of self-monitoring was Helmholtz (1866). His concern, however, was not with pathology, but with the following problem presented by healthy visual cognition. When an image moves across the retina, how does our brain know whether it is the world moving across our eyes or our eyes moving across the world? Helmholtz suggested that our brain can tell the difference because when our eyes move there is a motor command. More specifically, information about the motor command, which Sperry (1950) later dubbed the "corollary discharge", is used by the brain to predict the sensory consequences that would be produced by the eye movement. If the predicted and actual sensory consequences match then the brain infers that the change was self-generated and the conscious percept is adjusted accordingly. We can see exactly what happens when there is no such motor command, and hence no such adjustment, when we press on our eye with our finger. When we do this, the world itself seems to tilt and shake.

Frith and Done (1989) took delusions of control to arise as a result of this self-monitoring going wrong. In particular, there is a mismatch between the predicted and actual sensory consequences of the bodily movement and so (as with Helmholtz's ocular example) the movement is attributed to an external source. Whereas in Helmholtz's example, the recognition by the nervous system that a certain stimulus is self-produced causes a correction of the visual percept, in more typical bodily motor control, it results in sensory attenuation. The evolutionary benefit of this is clear enough: your nervous system needs to pay attention to stimuli that come from the outside, not the endogenous stimuli that (in a well-functioning system) will be harmless and irrelevant. Various data suggest that something goes wrong with this monitoring and subsequent attenuation in the context of schizophrenia (Frith et al. 2000). The most striking such datum is the reported finding that subjects with

diagnoses of schizophrenia can tickle themselves. The postulated explanation for this is that there is a mismatch between expected and actual sensory consequences and the sensory consequences are not attenuated: the tickling sensation is like being tickled by somebody else. Typical subjects can't tickle themselves because their nervous systems accurately monitor, and successfully attenuate, the sensory consequences of the tickling movements (Blakemore et al. 2000).

Whether the details of this account are accurate or not, what you get is a possible insight into both what the subject's experience might be like, and the underlying mechanistic abnormalities that may have given rise to this. In particular, due to faulty low-level self-monitoring, there isn't sensory attenuation, and therefore, even though the subject's bodily actions are successfully carried out, are in keeping with her intentions, they are experienced as strange.

We might call this switch from the mechanistic explanation of the experience, to the explanation of what is believed on the basis of what is experienced, "baton-passing". The explanation of the phenomenon is broken down into an answer to a subpersonal-level (A-type) question, and an answer to a personal-level (B-type) question. But there is no category mistake since these answers are not taken to answer the wrong questions. The mechanisms themselves are not taken to provide grounds for the subject's belief; the mechanisms explain the presence of a certain experience, and the experience provides the grounds. Although there is a personal-level explanation of the belief (viz. which appeals to the experience), there is no personal-level explanation of the experience itself, only a subpersonal one.[12] You cannot ask, "Why are you having this experience?" in the same way (viz. employing the same use of "Why") that you can ask, "Why did you do this?" or "Why do you believe this?"

Another nice example of how neuropsychiatry makes use of the personal/subpersonal distinction is in delusions of misidentification. Here not only do we see the distinction in use, but also we see how different hypotheses locate the "baton-passing" in a different place, namely, they differ about where personal-level explanation takes the baton (viz. is suitable).

In fact, it is interesting to see how something resembling Dennett's distinction was present in work on delusions in pre-Dennettian times. A key figure in the history of theoretical work on delusion is Karl Jaspers, who, in his *General Psychopathology* (1963), claimed that there were two very different projects in understanding mental illness. One involves "understanding the subject", and the other involves rendering the psychopathological phenomenon causally tractable. When it comes to delusions (in particular the primary delusions of schizophrenia), Jaspers claimed that they are "un-understandable", by which he meant that they could not be rendered intelligible in something very close to the sense that I have

[12] Analogously, the subpersonal psychology and neuroscience of early vision may explain to us how come certain visual illusions are experienced. But there is no personal-level explanation of this. There is only a personal-level explanation of why the illusion experienced is or is not taken at face value (i.e. leads to belief).

sketched here.[13] He suggested that since one of his two enterprises was impossible for these subjects, we should instead focus on trying to understand the subject as a causal system. This clearly resembles the claim that we should try to understand schizophrenia "subpersonally".

However, half a century later the way was paved for potentially rendering (at least some) delusions intelligible. In particular, Brendan Maher conjectured that the "delusional belief is not being held "in the face of evidence strong enough to destroy it," but is being held because evidence is strong enough to support it" (1974, p. 99). If this hypothesis is correct, then clearly, we need to understand what that evidence might be, and then we can answer the all-important personal-level question: "Why does the delusional subject believe what she does?" This experiential evidence will obviously have the potential to vary from one delusion to another, and presumably in a way that will be illuminating with regard to the nature and content of the delusion in question. In other words, there should be some clear connection between the nature of the experience and the content of the delusion. If this is the case, the nature of the experience tells us why one thing is believed and not another. Figuring out what this evidence (viz. the experience) might be, and how it arises, will likely require us to investigate at a subpersonal level.

This project received something of a breakthrough in the case of the Capgras delusion, the delusion that one or more loved ones have been replaced by identical-looking impostors. Borrowing Bauer's (1984) model for facial processing, whereby there are two streams for processing facial information – one covert, affective and anatomically dorsal, the other overt, semantic and anatomically ventral – Ellis and Young (1990) put forward the influential proposal that the Capgras delusion can be understood as a sort of "inverse prosopagnosia". Subjects with prosopagnosia have difficulty in the overt recognition of faces. Show them a picture of a familiar face and they will not be able to tell you whose face it is. And yet, surprisingly, some of them appear to have differential autonomic responses (roughly, affective/emotional responses) to these faces, as measured by heightened skin conductance response (SCR). In other words, although they themselves cannot tell you whose face they are looking at, their affective system seems at the very least to be able to "tell" that it is someone familiar. Ellis and Young hypothesized that Bauer's two streams can be selectively impaired, leading to double dissociation. According to them, whereas with prosopagnosia the affective stream for "covert recognition" is intact and the semantic stream for "overt recognition" is impaired, with the Capgras delusion it is the other way around. At a personal level this means that the Capgras patient is presented with someone who, thanks to intact semantic processing, looks to them exactly like a loved one, but there is a lack of affective response. The perceived person feels unfamiliar and the patient therefore concludes that this person cannot be the loved one in question. This model was given experimental support (Ellis et al.

[13] According to Jaspers they are not only formed differently from normal beliefs, but are also maintained differently: they are held with more conviction and are more tenacious in the face of contrary evidence.

1997) when it was discovered that, in contra-distinction to prosopagnosia, Capgras patients show diminished SCR when presented with familiar faces.

So there is at least scope for the Capgras delusion to be rendered intelligible, since it can be seen as something that is inferred on the basis of experiential evidence. Theories that take delusions to be grounded in unusual experiences are called "bottom-up" theories. Like Frith's account of delusions of control, a complete bottom-up account will contain a mix of personal and subpersonal explanation. However, this does not constitute the mixing that Dennett warns us against.[14] Again, here subpersonal answers aren't taken to answer personal-level questions. Rather, the nature of the question changes in accordance with our explanatory concerns, with what needs explaining. The presence of the anomalous experience is explained in terms of mechanism. In the Capgras case, on the Ellis and Young model, this involves explaining how lesions disrupt affective processing of familiar faces, and we get an explanatory connection between the damage and the presence of the experience on the one hand, and the nature of the experience and content of the delusion on the other.

Crucially the delusional judgment itself is explained at the personal-level, in terms of the quality of the experience. The relevant question to ask is: "*Why* does the *person* believe that this woman is not his mother?" And the relevant answer is something like: "Because this woman feels deeply unfamiliar to him." This is not a mechanistic explanation, but a personal-level one. It makes the belief intelligible. And, if correct, (to echo the early Dennett) it tells us all we need to know *within the scope of personal-level explanations*.

Note also that, although bottom-up theories explain delusions in terms of inferences on the basis of experiential evidence, this doesn't mean that these inferences should be thought of as rational. Indeed, the most widely accepted bottom-up theories hypothesize reasoning biases. Explaining the presence of such reasoning biases may require us to delve into subpersonal mechanisms once more. These so-called two-factor theories (e.g. Davies et al. 2001) claim that the bizarre experience is not enough to explain why the delusion is held for so long. The patient has a second deficit that renders the patient epistemically irrational. This is supported partly by non-empirical arguments, and partly by appeal to other patients who have similar affective deficits (the "first factor") towards loved ones, but who don't go on to form the delusion.[15] It is also supported by the fact that Capgras patients very often have two loci of damage, one temporal (or temporo-parietal) and another in the dorsolateral prefrontal cortex. It is the frontal damage that is taken to account for the bias.[16]

[14] That sort of mixing is, for example, answering "Why (in a grounds-seeking sense) does John believe that?" with reference to neurons and neurotransmitters.

[15] Although one might question whether the experience is exactly the same. SCR, which is what is appealed to here, can be disrupted in many different ways.

[16] The bias itself is then characterized in different ways by different theorists. For example, Kihlstrom and Hoyt (1988, p. 96) put it in terms of "non-optimal hypothesis-testing strategies".

Popular though this framework might be, not everyone subscribes to bottom-up theories of delusions (see, Campbell 2001). These theories want to place the "baton-passing" in a different place. In a way that harks back somewhat to Jaspers, these theorists claim that the delusion is not inferred, nor grounded in evidence, but merely caused. Any report (or even experimental evidence from SCR), for example, that the mother feels unfamiliar, is a consequence of (or an accompaniment to) the delusional belief, but not grounds for it. She feels unfamiliar because she is judged to not be the subject's mother, and not the other way around. An upshot of this is that the belief (or delusional state, if you don't want to call it a belief) can only be explained subpersonally. In answer to the same question "*Why* does the *person* believe that this woman is not their mother?" one cannot appeal to grounds or justification. One can only answer: "Because (in the justificatory sense of because) she just does." Again, in a way that echoes Jaspers, this unreasoned doxastic commitment is not only non-rationally *formed*; it is not open to rational *correction* either. This accounts for the so-called tenacity of delusions, but not, as bottom-up theories do, in terms of inference (biased or otherwise). An upshot of this is that, unlike with bottom-up theories, the delusion must be explained subpersonally. The only question with an illuminating answer is: "What has *caused* this person to act the way she does?" And again we are back to Jaspers' claim that delusional subjects are "un-understandable".

However, note that, although, on these top-down theories, the delusional belief itself may not be amenable to such an explanation, any *action* (arguably by definition, if it really is an action) performed on the basis of the belief will be amenable to such an explanation, and this explanation will appeal to the belief. In such a situation we get something like the following series of questions and answers.

Q: Why did the patient stab her father (even though they seemed to have a good relationship prior to the event)?
A: Because she believed that he was not her father, but an identical-looking impostor.
Q: And on what grounds did she believe this?
A: We can't say. She just did.

At this point we would need to delve into the subpersonal mechanisms to understand what is underpinning the (unintelligible) belief mechanistically.

Both top-down and bottom-up theories implicitly make use of a personal/subpersonal distinction, and they are making use of the same distinction. Where they differ is about substantive, empirical facts about what is going on inside these patients, and, as a result, they locate the baton-passing from subpersonal to personal-level explanation in a different place.

6.5 Troubled Persons or Broken Intentional Systems?

Many people with mental disorders, as a consequence of these disorders, behave in strange and unpredictable ways. Delusional subjects, for example, don't believe what you'd expect them to, given their apparent epistemic or informational situation,

and, by extension, they act strangely and make unexpected claims. Not only this, they sometimes fail to act in accordance with their claims. For example, although Capgras patients claim that their loved ones have been replaced by impostors, they often fail to worry about the welfare or whereabouts of the replaced loved one.[17]

Recall that the later version of the personal/subpersonal distinction, which is tied to the intentional stance, suggests that something is an intentional system to the extent that it can be predicted by the intentional stance, namely, by ascriptions of mental states (e.g. beliefs and desires). It seems as though such a view would have to claim that these patients either aren't intentional systems at all, or perhaps that they are severely deficient ones that are extremely hard to predict. Remember that "first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have given its place in the world and its purpose [etc.]" (1981, p. 558). In the Capgras case, the patient ought not (in the sense that we would not expect her) to have the delusional belief. Indeed, that is arguably one of the reasons (perhaps *the* reason) why she is branded delusional. Perhaps we will want to say that once she makes the delusional assertion, it seems that we have grounds, on the intentional stance, for attributing that belief to her. But then she doesn't act in accordance with this assertion, so we then have grounds for retracting the attribution.[18] The intentional stance runs into similar difficulties with a plethora of other mental disorders.

Is this subject less of an intentional system because she is so hard to predict? It seems that given what Dennett says about the intentional stance, and about the way that personal-level mental state attributions are merely accurate to the extent that successful behavioral prediction can be achieved using the stance, we have to bite the bullet here. Perhaps we might even say that these people, through their inconsistency and unpredictability, are more defective intentional systems than, say, a well-functioning chess computer.

However, it seems like we can, in principle, give personal-level explanations of irrational and unpredictable behavior in human subjects. Indeed, I've been suggesting that neuropsychiatry as currently practiced attempts, where possible, to do precisely this. Take yet another example. Addictive behavior (e.g. compulsive gambling or drug-seeking behavior) is deeply irrational and hard to predict using the intentional stance. And yet we can give explanations of addictive behavior in terms of personal-level urges, the presence of which can be explained in terms of a subpersonal hijacking of the reward-system. Although strong urges (a bit like pains) lead to irrational behavior, they can be appealed to in giving a personal-level explanation that renders the subject intelligible. The same sorts of things cannot be said of a malfunctioning chess computer. When a malfunctioning chess computer makes a

[17] You also get this inertia in other delusions, including schizophrenia. As the psychiatrist Bleuler pointed out, "none of our generals has ever attempted to act in accordance with his imaginary rank and station" (Bleuler 1950).

[18] Indeed there is an important philosophical debate about whether delusions are beliefs, namely, whether patients really believe what they sincerely assert. Bayne and Pacherie (2005), for example, say "Yes". Currie (2000) and Currie and Jureidini (2001), say "No" (They merely believe that they believe it, and in fact only imagine it).

perplexingly poor move, there is no "personal-level" fact that can be appealed to in order to explain this. The intentional stance helps us *predict* the well-functioning chess computer by enabling us to see the relevant patterns. But when the chess computer malfunctions, it is so different from us that we would never ask to render its malfunctioning behavior *intelligible*, let alone expect to be able to do so.

On Dennett's earlier, more Rylean view, personal-level explanations aren't predictions churned out by adopting the intentional stance. They are satisfying answers to particular kinds of question, which acquire their significance from the use we make of personal-level vocabulary in true and meaningful everyday discourse. This allows that irrational and unpredictable behavior can be rendered intelligible if we understand what the subject is going through (e.g. urges, pains, feelings of unfamiliarity). This is not to say that personal-level explanations will always be available (e.g. if top-down theorists are right, then they won't be available for explaining the Capgras delusion), but rather that, when they are available, giving them is illuminating, and failing to give them is missing something of vital importance.

6.6 Conclusion

To sum up, then, the personal/subpersonal distinction is extremely important for accounts that help us to understand subjects with mental disorders. It was introduced by Dennett in order to prevent category mistakes, and, in particular, to prevent people from trying to provide mechanistic, subpersonal-level answers to non-mechanistic personal-level questions. When Dennett later introduced the intentional stance, his aim was rather different, and the personal-level became simply that which enables us to predict behavior when we are using the intentional stance.

We looked at the implicit use that cognitive neuropsychiatry already makes of the personal/subpersonal distinction. We then noted that the version of the distinction that is in operation seems closer to the earlier than the later one, since personal-level explanations are often given of irrational and unpredictable behavior.

This is obviously not a knock-down argument against the later version of the distinction, but it isn't supposed to be. Rather, I hoped to clarify both versions of the distinction (and in so doing, one aspect of the evolution of Dennett's thinking) by contrasting them with one another. I also hoped to show the importance of the distinction in general, and the early version in particular, for the important and exciting field of cognitive neuropsychiatry.

References

Bauer, R. M. (1984). Autonomic recognition of names and faces in prosopagnosia: A neuropsychological application of the guilty knowledge test. *Neuropsychologia, 22*(4), 457–469.

Bayne, T., & Pacherie, E. (2005). In defence of the doxastic conception of delusion. *Mind & Language, 20*(2), 163–188.

Blakemore, S., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself? *NeuroReport, 11*(11), R11–R16.
Bleuler, E. (1950). *Dementia praecox or the group of schizophrenias*. New York: Trans. Joseph Zinkin, International Universities Press.
Campbell, J. (2001). Rationality, meaning and the analysis of delusion. *Philosophy, Psychiatry, and Psychology, 8*(2–3), 89–100.
Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind/brain*. Cambridge, MA: MIT Press.
Currie, G. (2000). Imagination, delusion and hallucinations. In M. Coltheart & M. Davies (Eds.), *Pathologies of belief* (pp. 167–182). Oxford: Basil Blackwell & Mott, Ltd.
Currie, G., & Jureidini, J. (2001). Delusions, rationality, empathy. *Philosophy, Psychiatry and Psychology, 8*(2–3), 159–162.
Davies, M., Coltheart, M., Langdon, R., & Breen, N. (2001). Monothematic delusions: Towards a two- factor account. *Philosophy, Psychiatry and Psychology, 8*(2/3), 133–158.
Dennett, D. C. (1969). *Content and consciousness*. London: Routledge.
Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy, 68*, 87–106.
Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT University Press.
Dennett, D.C. (1981). True Believers: The intentional strategy and why it works. In A. F. Heath (Ed.), *Scientific explanation*: Papers based on Herbert Spencer lectures given in the University of Oxford. Clarendon. Reprinted in D. Chalmers (Ed.) (2002). *Philosophy of mind: Classical and contemporary readings*. Oxford University Press.
Ellis, H. D., & Young, A. W. (1990). Accounting for delusional misidentifications. *British Journal of Psychiatry, 157*, 239–248.
Ellis, H. D., Young, A. W., Quayle, A. H., & de Pauw, K. W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society of London: Biological Sciences, B264,* 1085–1092
Frith, C. (1992). *The cognitive neuropsychology of schizophrenia*. Hove: Lawrence Erlbaum.
Frith, C., & Done, D. (1989). Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychological Medicine, 19*, 359–363.
Frith, C., Blakemore, S.-J., & Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews, 31*(2–3), 357–363.
Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge, MA: MIT Press.
Helmholtz, H. von (1866). Concerning the perceptions in general. In *Treatise on physiological optics*, vol. III, 3rd edn (translated by J. P. C. Southall 1925 Opt. Soc. Am. Section 26, reprinted New York: Dover, 1962).
Jaspers, K. 1963. *General psychopathology* (trans: Hoenig, J., & Hamilton, M.). Manchester: Manchester University Press.
Kihlstrom, J. F., & Hoyt, I. P. (1988). Hypnosis and the psychology of delusions. In T. F. Oltmanns & B. A. Maher (Eds.), *Delusional beliefs: Interdisciplinary perspectives* (pp. 66–109). New York: Wiley.
Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology, 30*, 98–113.
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283–311.
Ryle, G. (1949). *The concept of mind*. Chicago: University of Chicago Press.
Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology, 43*(6), 482–489.
Wilkinson, S. (2014). Levels and kinds of explanation: Lessons from neuropsychiatry. *Frontiers in Psychology, 5*, 373.

Chapter 7
I Am Large, I Contain Multitudes: The Personal, the Sub-personal, and the Extended

Martin Roth

Abstract In *Content and Consciousness*, Daniel Dennett introduces a distinction between personal and sub-personal levels of explanation. Minding the distinction is key to avoiding false starts and dead ends, Dennett warns, especially when it comes to the areas of thinking and reasoning. Why is the distinction important? To what extent have cognitive scientists and philosophers honored this distinction? This paper will use the current debate over the extended mind hypothesis – roughly, the claim that 'mental' or 'cognitive' processes extend beyond the boundaries of the brain – to approach both questions. There are several reasons why investigating the extended mind debate is apt: not only has it garnered the attention of some of our most creative and important researchers in cognitive science, but as will be shown, lurking behind the debate are largely unacknowledged assumptions about how and why the personal/sub-personal distinction should be drawn. To show this, the paper will first look at some key differences in how Dennett and Jerry Fodor interpreted Gilbert Ryle and the way those differences showed up in their respective treatments of the personal/sub-personal distinction. The paper will then consider – and provide a partial defense of – a version of the extended mind hypothesis that honors the personal/sub-personal distinction. Finally, the paper will survey some of the recent literature on the extended mind hypothesis and argue that several of the ways the hypothesis has been discussed display the very confusions that Dennett warns us against.

7.1 Introduction

In the spirit of Kant, we can read Daniel Dennett's *Content and Consciousness* (1969; *C&C*, henceforth) as an attempt to answer the question "How is a science of the mind possible?" Dennett wasn't the only philosopher exploring this territory at

M. Roth (✉)
Drake University, Des Moines, IA, USA
e-mail: martin.roth@drake.edu

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_7

the time, of course, and as attempts to articulate a conceptually coherent foundation for a science of mind go, Jerry Fodor's work was equally pioneering. There were important differences, however, in the ways Dennett and Fodor conceived of the *constraints* under which such a science could succeed, and a chief reason for those differences was a difference in how Dennett and Fodor responded to Ryle; whereas Fodor thought that the possibility of a science of the mind required refuting Ryle, Dennett's project was to show how it was possible to have a robust science while still honoring Ryle's insights. For Dennett, a key move is the distinction between personal and sub-personal levels of explanation. Dennett introduces the distinction at the end of Part I, and he treats this section as an opportunity to "consolidate the gains" of the first part of the book. One of the issues at stake here is whether explanations that appeal to intentionally characterized states and processes have a substantive role to play in science, and Dennett's answer is that although such explanations can play a substantive role in science, many of the explanations we can expect to find will not – indeed, *cannot* – appeal to intentionally characterized states and processes applicable to whole persons. Minding the distinction is key to avoiding false starts and dead ends, Dennett warns, especially when it comes to the areas of thinking and reasoning (p. 167).

Why is the distinction between personal and sub-personal levels of explanation important? To what extent have cognitive scientists and philosophers honored this distinction? This paper will use the current debate over the extended mind hypothesis – roughly, the claim that 'mental' or 'cognitive' processes extend beyond the boundaries of the brain – to approach both questions. There are several reasons why investigating the extended mind debate is apt: not only has it garnered the attention of some of our most creative and important researchers in cognitive science, but as will be shown, lurking behind the debate are largely unacknowledged assumptions about how and why the personal/sub-personal distinction should be drawn. To show this, the paper will first look at some key differences in how Dennett and Fodor interpreted Ryle and the way those differences showed up in their respective treatments of the personal/sub-personal distinction. The paper will then consider – and provide a partial defense of – a version of the extended mind hypothesis that honors the personal/sub-personal distinction. Finally, the paper will survey some of the recent literature on the extended mind hypothesis and argue that several of the ways the hypothesis has been discussed display the very confusions that Dennett warns us against.

7.2 The Legacy of Ryle and the Birth of the Personal/Sub-personal Distinction[1]

Some people know how to multiply, and among those who do, some occasionally display their knowledge on paper. How should we understand the relationship between knowing how to multiply and such particular displays of this knowledge?

[1] Some of the points discussed in this section are developed in Cummins and Roth (2011).

In Chap. II of *The Concept of Mind*, Gilbert Ryle characterizes and critiques an *intellectualist* answer to this kind of question. According to the intellectualist, knowing how to multiply consists in propositional knowledge (knowledge-*that*) of ways to multiply. When people act on the intention to solve a multiplication problem, they first consult the relevant propositional knowledge and formulate a strategy for applying this knowledge.[2] The calculations performed are outputs of this planning process. Generalizing, if we think of propositional knowledge of ways to X as a "mini-theory" of how to X, then we can say that actions that manifest know-how are the causal consequences of a theory-driven planning process. As Ryle put it, we first do a bit of theory and then do a bit of practice (1949, p. 29).[3] Famously, Ryle argued that intellectualism leads to an unacceptable regress: since theorizing itself is something we know how to do, explaining particular instances of theorizing would require positing still *further* acts of theorizing, ad infinitum (p. 30). Hence, knowing-how cannot be defined in terms of knowing-that (p. 32). In what does knowing-how consist, then? According to Ryle, know-how was to be found in "capacities, skills, habits, liabilities, and bents" (p. 45). Exercises of know-how are simply manifestations of capacities, where capacities are understood by Ryle to be dispositions "the exercises of which are indefinitely heterogeneous" (p. 44).

As Dennett understood him, Ryle was doing philosophy, not science. Thus, while Ryle's identification of know-how with dispositions may have made him some sort of "philosophical behaviorist," Dennett did not take Ryle's argument to entail Skinnerian psychology.[4] The difference between Fodor and Dennett on this point is crucial. Fodor and Dennett agreed about the inadequacies of Skinnerian psychology, and each recognized the importance of Hilary Putnam's machine functionalism (Putnam 1960, 1967) to the development of a cognitive science rich with intentional characterization. As we will see, however, Fodor and Dennett *disagreed* over the significance of the personal/sub-personal distinction, and at the root of the disagreement was a disagreement about how to interpret Ryle. According to Fodor, Ryle's attack on *mentalism* – roughly, the view that behavior can be explained in term of causally efficacious mental states and processes – rested on the assumption that mentalism entails dualism. Because dualism was not a serious option, and because Ryle thought that the only alternative to dualism was behaviorism, Ryle was a behaviorist. This reading of Ryle is indicated clearly in *Psychological Explanation* when, after laying out the functionalist alternative to behaviorism, Fodor writes, "[O]nce it has been made clear that the choice between dualism and behaviorism is not exhaustive, a major motivation for the defense of behaviorism is removed: we are not required to be behaviorists simply in order to avoid being dualists" (1968a, p. 59). Fodor repeats the point in *The Language of Thought*: "Ryle assumes … that a mentalist must be a dualist; in particular, that mentalism and materialism are mutually exclusive" (1975, p. 4). Note that if we read Ryle this way,

[2] In this case, the relevant propositional knowledge would be something like knowledge of instructions or procedures for doing multiplication, e.g., an algorithm.

[3] All Ryle references are to *The Concept of Mind*.

[4] Dennett (1978) is especially clear on this point.

it will be tempting to conclude that we can salvage intellectualism by embracing functionalism; after all, intellectualism implies mentalism, so if Ryle banished mentalism because mentalism entails dualism, then it will make sense to think that Ryle's chief objection to intellectualism was that it leads to dualism.

Of course, Fodor *did* try to salvage intellectualism. In "The Appeal to Tacit Knowledge in Psychological Explanation," Fodor defends intellectualism as an account of mental competences, which Fodor identifies with abilities, e.g., the ability to play chess, the ability to type 'Afghanistan,' and the ability to speak Latin (1968b, p. 72). To give a feel for his version of intellectualism, Fodor begins his paper with the following explanation of how we tie our shoes:

> There is a little man who lives in one's head. The little man keeps a library.
> When one acts upon the intention to tie one's shoes, the little man fetches down a volume entitled *Tying One's Shoes*. The volume says such things as: "Take the left free end of the shoelace in the left hand. Cross the left free end of the shoelace over the right free end of the shoelace…," etc. When the little man reads the instruction 'take the left free end of the shoelace in the left hand,' he pushes a button on a control panel. The button is marked 'take the left free end of a shoelace in the left hand.' When depressed, it activates a series of wheels, cogs, levers, and hydraulic mechanisms. As a causal consequence of the functioning of these mechanisms, one's left hand comes to seize the appropriate end of the shoelace. Similarly, mutatis mutandis, for the rest of the instructions. The instructions end with the word 'end.' When the little man reads the word 'end,' he returns the book of instructions to his library. That is the way we tie our shoes. (pp. 63–64)

According to this explanation, shoe tying behavior is mediated by internal processes that employ "propositions, maxims, or instructions" (p. 76) regarding shoe tying, and Fodor urges us to take seriously the hypothesis that a whole host of abilities are best explained in this way. Moreover, Fodor argues that this hypothesis yields an intellectualist account of know-how: knowing how to do something consists in having an ability whose exercises have the right kind of *etiology*: "If an organism knows how to X, then nothing is a simulation of the behavior of the organism which fails to provide an answer to the question 'How do you X?'" (p. 75). In other words, the ability to X constitutes knowing how to X only if the causal processes that generate X-ing behavior represent a way to X. In the shoe-tying example, the person *knows how* to tie shoes since a representation of a way to tie shoes is part of the process that generates shoe-tying behavior. However, the sorts of propositions, maxims, or instructions that are involved in this story about shoe tying need not be consciously available to the person (in the minimal sense that the person need not be able to report them). In such cases we can say that the person has *tacit* knowledge of these propositions.

Fodor and Dennett recognized that a psychological theory is complete only if it makes no reference to unanalyzed psychological processes, and Fodor acknowledges that the little man is merely a temporary stand-in for whichever psychological faculties turn out to apply the information about how to tie shoes (p. 65). In this way, Fodor and Dennett each accepted a version of *homuncular functionalism* (Dennett 1978; Lycan 1987). Homuncular functionalism is a particular instance of functional analysis (Cummins 1975) in which the complex capacities of a person are analyzed into simpler capacities of sub-personal systems. The capacities of the sub-personal

systems are in turn analyzed into even simpler capacities until we reach capacities whose exercise requires no know-how. However, Fodor and Dennett disagreed over what the implications of drawing the personal/sub-personal distinction *were*. For reasons we will see in a moment, Dennett thinks the personal/sub-personal distinction *blocks* the move from 'the little man employs rules' to 'the person employs rules.' By contrast, Fodor claims that "[T]he intellectualist account of X-ing says that, whenever you X, the little man in your head has access to and employs a manual on X-ing; and surely, whatever is his is yours" (pp. 73–74). This, in turn, permits Fodor to say that tacit knowledge of how to tie shoes is knowledge that a *person* has, and it is the *person* who employs this knowledge in the production of behavior:

> What, then, are we to say is the epistemic relation an agent necessarily bears to rules he regularly employs in the integration of behavior? There is a classic intellectualist suggestion: if an agent regularly employs rules in the integration of behavior, then if the agent is unable to report these rules, then it is necessarily true that the agent has *tacit* knowledge of them. (p. 74)

Putting aside for the moment whether Fodor is entitled to this conclusion, note that Fodor did not think the inference resulted from an oversight on his part. The following passage from *The Language of Thought* makes this clear:

> There is, obviously, a horribly difficult problem about what determines what a person (as distinct from his body, or parts of his body) did. Many philosophers care terrifically about drawing this distinction … [B]ut whatever relevance the distinction between states of the organism and states of its nervous system may have for *some* purposes, there is no reason to suppose that it is relevant to the purposes of cognitive psychology. (p. 52)

As Fodor well knew, Dennett was among the many philosophers who cared terrifically about drawing this distinction. But this leaves us with a puzzle: if Fodor and Dennett each subscribed to homuncular functionalism, why did they disagree over the importance of the personal/sub-personal distinction to cognitive psychology? To answer this, we need a close look at how and why Dennett drew the personal/sub-personal distinction.

Dennett introduces the personal/sub-personal distinction with the example of pain. People can distinguish pains from other sensations. How do they do this? If the question is asking about activities people perform, then the question does not admit of an answer: a person does not do anything in order to distinguish pains (*C&C*, p. 103). This does not mean that the ability to distinguish pains admits of no explanation, but in attempting to give such an explanation "we must abandon the explanatory level of people and their sensations and activities and turn to the *sub-personal* level of brains and events in the nervous system. But when we abandon the personal level in a very real sense we abandon the subject matter of pains as well" (*C&C*, p. 105). Now, abilities constituting know-how (Fodor's "mental competences") are among the abilities we want to explain, and while Dennett acknowledged that some abilities can be analyzed in terms of further personal activity, e.g., discriminating good apples from bad may be analyzed into checking for color and crispness (*C&C*, p. 104), at a certain point a 'mental process' story told at the

personal level will no longer be available. There may be something *like* a mental process story available at the sub-personal level, but just as the sub-personal *explanation* of pain must not include talk of pains, sub-personal explanations of intellectual activity, e.g., thinking and reasoning, ultimately must drop talk of intellectual activity in the analysis. Here is how Dennett puts the point:

> People *arrive at* conclusions, and, as the bland verb suggests, this is not a process that people go through or an activity in which they engage, so we cannot ask the question 'How do you arrive at a conclusion?' and expect an answer in the form 'First I do this, and then I do that'; people do not do anything in order to arrive at conclusions, but their brains must. The distinction between personal and sub-personal levels of explanation is nowhere more important than in the area of thinking and reasoning … Were we to take what goes on in the brain and analyse it into parts, we should not expect those parts to be, say, concluding or deducing operations, for that is to confuse levels, and yet some operations of a different sort must occur. When computers are made to perform logical operations, the abstract, timeless transformations and operations of logic are realized in physical, temporal operations, and the production of results or conclusions takes time and energy. (*C&C*, p. 167)

If we appeal to operations performed by brain parts in an effort to *explain* how people arrive at conclusions, then the resulting explanation, given in terms of operations performed by brain parts, will be sub-personal. But sub-personal explanations cannot include terms used to describe the activities of people, for the resulting explanations would be *empty*. Putting matters this way helps to illuminate why Dennett finds Fodor's explanation of shoe tying objectionable. Dennett puts the objection as follows:

> In his purest form the little man in the brain takes on the guise of brain-writing reader, an intelligent, communicating system capable of understanding messages. Positing the brain-writing reader is almost irresistible, for if we cannot understand central states and events of the nervous system as bearing content, as being messages of some sort, it is not clear how we can understand them at all. The temptation must be resisted, however, by recognizing the disanalogies between verbal communication and non-verbal intra-cerebral communication and indeed the primacy of non-verbal communication. Other roles played by the little man in the brain are merely specialized roles projected inwards from the details of our initial analysanda, the variety of affairs of a person. The solitary audience in the theatre of consciousness, the internal decision-maker and source of volitions or directives, the reasoner, if taken as parts of a person, serve only to postpone analysis. The banishment of these concepts from our analysis forces the banishment as well of a variety of other self-defeating props, such as the brain-writing to be read, the mental images to be seen, the volitions to be ordered, and the facts to be known. These props are self-defeating because they could only serve the functions for which they were designed in conjunction with interior person-analogues, and hence as elements in an analysis they reproduce the problems like images in a hall of mirrors. (*C&C*, pp. 214–215)

The problem with Fodor's explanation of shoe tying is not that it invokes intentionally characterized states and processes; indeed, a major aim of *C&C* is to defend such characterizations.[5] The problem is that the terms used to describe the little

[5] Dennett articulates and defends a "teleofunctional" account of content, and while a close examination of the details of this account is necessary for a full understanding of the vision of cognitive science provided by *C&C*, a close examination of those details is not required here. For present purposes, the constraints that the personal/sub-personal distinction places on explanations invok-

man's activities (e.g., reading, writing, and sending messages) are terms we use to describe the activities of language users like us – people – and while Dennett acknowledges that sub-personal processes may be somewhat like the activities of people, we need to resist the temptation to think that the contents of sub-personal states can be associated with verbal expressions of a natural language: "Verbal expressions, however, are not the ultimate vehicles of meaning, for they have meaning only in so far as they are the ploys of ultimately non-linguistic systems. The inability to find precisely worded *messages* for neural vehicles to carry is thus merely an inability to map the fundamental on to the derived, and as such should not upset us" (p. 99). The relevance to explanation is this: even if a mapping existed, providing the mapping would not amount to *explaining* personal level activities; at best, it would provide us with a way to *redescribe* those personal level activities. That Fodor may be conflating explanation with redescription is suggested by the good deal of space he dedicates to showing that, "[A] programming language can be thought of as establishing a mapping of the physical states of a machine onto sentences of English such that the English sentence assigned to a given state expresses the instruction the machine is said to be executing when it is in that state" ("Tacit," p. 76). For Fodor, establishing these mappings is crucial to the debate over intellectualism because establishing these mappings is crucial to the defense of tacit knowledge attributions. But if you do not think the debate over intellectualism turns on whether intellectualism can be made metaphysically respectable, the effort Fodor expends to dress up tacit knowledge in computational clothing will be beside the point. The problem with intellectualism is not that it is metaphysically suspect. The problem is that it attempts to explain know-how in terms of personal level activities. It is not enough to get rid of the little man; we also need to throw out his book.[6]

Throwing out the book would not signal a return to Skinnerian behaviorism, however. One of the main aims of *C&C* is to show that the limitations imposed by Ryle apply only to the *personal* level of explanation (p. 107), and while acknowledging this limitation is consistent with a cognitive science rich with intentional characterization, we must avoid the attempt to found cognitive psychology on an account of content that is tailor-made for ascriptions of content at the personal level. What cognitive psychology needs is an understanding of how information can be acquired, stored and manipulated in a way that gives rise to intelligent and adaptive behavior, including, in the case of humans, and perhaps some other creatures, the ability to understand language. There *is* a kind of content in the brains of intelligent creatures, but it isn't content, as that is generally understood. Linguistic content is

---

ing intentionally characterized states and processes matter most, and those constraints turn more on the demands of explanation than they do on the particular shape Dennett's theory of content takes.

[6] As the passage quoted on page 8 suggests, Dennett isn't claiming that neural events and processes cannot be described in terms of computations. In fact, *C&C* contains a lengthy discussion of how we might provide such descriptions (Chap. 5). Dennett's point is that even if such descriptions were key to justifying ascriptions of content, it would not follow that the contents ascribed are contents ascribable to persons. More to the point, such ascriptions *cannot* apply to persons if such ascriptions are part of what is required to *explain* personal level phenomena.

an *explanandum* for cognitive science, not an *explanans*, and for those worried about the pitfalls of trying to read off the fundamentals of cognition from the linguistic capacities of people, getting the semantics of natural language right should probably be pretty far down on the agenda. And even there, one must beware of the kind of translation theory of language understanding that makes the language of thought hypothesis seem inevitable.[7] This simply follows from what Dennett sees the *point* of sub-personal explanations to be: to explain those personal-level capacities that have no further explanation at the personal level. Sub-personal intentional characterizations provide an explanatory bridge between the intentional characterizations we use to describe people and the physical language we use to describe brains and bodies.[8]

7.3 How to Make Yourself Large

> It is part of the function of this book to show that exercises of qualities of mind do not, save *per accidens*, take place 'in the head', in the ordinary sense of the phrase, and those which do have no special priority over those which do not. (Ryle, p. 40)

Ryle's attack on intellectualism was not an attack on mental processes. For example, some people know how to do long division, and doing long division is a mental process (p. 22). Furthermore, Ryle did not deny that some mental processes take place in the head. Rather, what Ryle tried to show is that the issue of whether a process is *inner or outer* is orthogonal to the issue of whether a process is *mental or non-mental*. Consider multiplication. My ability to multiply 41×17 analyzes into my ability to multiply 7×1, add 4 and 2, and so forth, and if exercises of the analyzing abilities are orchestrated in the right way,[9] the result is the exercise of the ability to multiply 41×17. But notice that this explanation-by-analysis tells us nothing about *where* exercises of these abilities take place; as far as the explanation goes, the exercises can take place in the head or they can take place on paper.

---

[7] The dangers of conflating meaning and content are explored further in Cummins and Roth (2012).

[8] The following passage suggests that such a bridge is required: "Since we cannot very well claim to have explained a mental phenomenon if we are unable to say (in the scientific language of our explanation) when a sentence heralding the occurrence of the phenomenon is true and when not, our task will involve at least this much: framing within the scientific language the criteria – the necessary and sufficient conditions – for the truth of mental language sentences" (p. 21). However, Dennett accepts that there are perfectly legitimate personal-level descriptions, e.g., descriptions in terms of thinking and reasoning. Nothing Dennett says indicates that those kinds of descriptions are incomplete (as personal level descriptions), and he does not urge that they be replaced by, or reduced to, sub-personal descriptions. In this way, the personal-level enjoys a kind of autonomy from the sub-personal level. These points apply, mutatis mutandis, to the relationship between sub-personal descriptions and descriptions couched in the language of physical science. The explanatory bridge thus should not be thought of as a reductive bridge.

[9] Orchestrated in a way that can be specified by an algorithm, e.g., a partial products algorithm.

In claiming that processes taking place in the head do not have a special priority over those taking place on paper, we need not deny that *neural* processes have a special priority over those that take place on paper. This point is crucial, but it is likely to be overlooked or misinterpreted if we fail to honor the distinction between personal and sub-personal levels of explanation. My ability to multiply 41×17 analyzes into my ability to multiply 7×1, add 4 and 2, and so forth, and because the analyzing abilities are *my* abilities, the analysis amounts to a *personal* level explanation of my ability to multiply 41×17. But if we turn to the analyzing abilities themselves, the personal level explanation drops out. For example, if the question is "How do you add 4 and 2?" the answer is that I don't do anything else; I just add them. At the *personal* level, my ability to add 4 and 2 amounts to little more than reliably responding with the correct answer when the problem is posed, and it is at the *personal* level where abilities exercised in the head have no special priority over abilities exercised on paper.

However, insofar as my ability to add 4 and 2 depends on *sub-personal* states and processes, and neural states and processes realize sub-personal states and processes, neural states and processes *do* have a special priority over abilities exercised on paper: the former *enable* the latter. But this point holds just as well for personal level abilities *that are exercised in the head*. To put the point another way: the priority that neural states and processes have over abilities exercised on paper has nothing to do with the distinction between inner and outer, but rather the distinction between sub-personal and personal.

Adopting this perspective opens the door to a rapprochement of two otherwise seemingly incompatible claims: (a) problem solving, calculating, figuring things out – these are all cognitive processes, and lot of these processes take place outside of the head, e.g., on scratch paper, iPads, whiteboards, and so on; (b) problem solving, calculating, figuring things out – these processes are made possible by cognitive processes that occur solely within the confines of the skull. The key to reconciliation, of course, is to note that the cognitive processes mentioned in (a) belong to the personal level, while the cognitive processes mentioned in (b) belong to the sub-personal level.

Whatever the prospects of the aforementioned perspective, a cursory look at the literature reveals that the way the debate over the extended mind hypothesis is typically framed promises to tell us very little about those prospects. The reason, in short, is that the participants to the debate have largely sided with Fodor on the issue of whether the distinction between what a person does and what a person's parts do is important to cognitive psychology. This isn't to say that the alignment has been acknowledged, i.e., this isn't to say that the personal/sub-personal distinction has been openly discussed and its importance found wanting. Indeed, there has been a conspicuous absence of such a discussion. When combined with an examination of the way positions and arguments have been formulated, it appears that either the personal/subpersonal distinction is assumed to be so obviously irrelevant that it does not merit a mention or it has simply not occurred to anyone that the distinction might be relevant to the debate over the extended mind hypothesis.

For instance, consider the way the debate has been framed by Frederick Adams and Kenneth Aizawa (prominent critics of the hypothesis) and Andy Clark (a prominent supporter of the hypothesis). In the course of characterizing the extended mind

hypothesis, Adams and Aizawa write, "It's certainly a wild idea to suppose that to use a calculator is to have one's mind bleed out of one's brain into plastic buttons and semiconductors" (2001, p. 44). Here is the proposal Adams and Aizawa intend to defend, a proposal that they think is antithetical to the extended mind hypothesis: "In this paper, we propose to defend common sense. Our view is that, as a matter of contingent empirical fact, in all actual cases of human tool use brain-bound cognitive processes interact with non-cognitive processes in the extracranial world" (2001, p. 46). Clark calls the thesis Adams and Aizawa defend BRAINBOUND, the thesis according to which

> …all human cognition depends directly on neural activity alone. The neural activity itself may, of course, in turn depend on worldly inputs and gross bodily activity. But that would be merely what Hurley usefully dubs 'instrumental dependence,' as when we move our head or eyes and get a new perceptual input as a result. All that really matters as far as the actual mechanisms of human cognition are concerned, BRAINBOUND asserts, is what goes on in the brain. (2011, p. xxvii)

By contrast, Clark defends a thesis he calls EXTENDED. EXTENDED

> …is a view according to which thinking and cognizing may (at times) depend directly and noninstrumentally upon the ongoing work of the body and/or the extraorganismic environment … According to EXTENDED, the actual local operations that realize certain forms of human cognizing include inextricable tangles of feedback, feed-forward, and feed-around loops: loops that promiscuously criss-cross the boundaries of brain, body, and world. The local mechanisms of mind, if this is correct, are not all in the head. Cognition leaks out into the body and world. (Clark 2011, p. xxviii)

To illustrate the difference between BRAINBOUND and EXTENDED, we can turn again to the example of multiplying large numbers on paper. Clark writes:

> The brain learns to make the most of its capacity for simple pattern completion (4 X 4 = 16, 2 X 7 = 14, etc.) by acting in concert with pen and paper, storing the intermediary results outside the brain, then repeating the simple pattern completion process until the larger problem is solved. The brain thus dovetails its operation to the external symbolic resource. The reliable presence of such resources may become so deeply factored in that the biological brain alone is rendered unable to do the larger sums. (2003, p. 6)

As suggested by the last line, if the brain alone is unable to do the larger sums, then it must be the brain *plus* something else (e.g. pen and paper) that does the larger sums. Since doing sums is clearly cognitive, the brain, pen, and paper are part of a cognitive system. It is in this sense that the local mechanisms of mind are not all in the head. By contrast, while Adams and Aizawa might agree that the brain cannot do certain sums without the presence of pen and paper, they do not think it follows that pen and paper are part of a cognitive system that solves the problem. The pen and paper enable the occurrence of the cognitive processes that are required to solve the problem, but as far as the cognitive processes themselves go, those processes take place in the brain exclusively.[10] In fact, there is a tension in Clark's own

[10] To infer that the pen and paper are part of a cognitive process from the fact that pen and paper enable a cognitive process is to commit what Adams and Aizawa call the "coupling-constitution fallacy" (2010, p. 91).

characterization of multiplying large numbers, one that appears to actually lend support to Adams and Aizawa's position. Clark's talk of "storing the intermediary results outside the brain" suggests that it is the *brain* that is performing the calculations. The results stored on the page thus look merely to be *outputs* of one calculation that can then serve as *inputs* to another calculation. If the relevant cognitive activity here is the calculating, then it appears that the paper and pen play no role in the calculating itself. But if this is correct, it suggests that the dependence on pen and paper is merely instrumental. And if *that* is correct, then it looks like BRAINBOUND has it right.

But BRAINBOUND doesn't have it right, even if Clark's formulation is subject to the aforementioned response. The problem is with Clark's formulation of EXTENDED. If we grant that the relevant cognitive activity is the calculating, then what Clark *should* say is that the calculating takes place on paper. My ability to multiply 41×17 analyzes into my abilities to multiply 7×1, add 4 and 2, and so forth, and to perform a calculation is to exercise the analyzing abilities in some orchestrated way.[11] In this case, the analyzing abilities are exercised on paper, and thus so is the calculating. If we interpret BRAINBOUND as a claim about *sub-personal* processes, however, then, far from denying it, defenders of EXTENDED can say this is precisely how we should understand how certain *personal* level capacities are possible. Brain-bound cognitive processes are sub-personal cognitive processes instantiated in brains. The paper and pencil existing outside of the skull are not part of an extra-cranial cognitive process that interacts with a brain-bound cognitive process. Rather, it is the interaction between brain-bound, *sub-personal cognitive processes* and external, *non-cognitive* tools that gives rise to exercises of *personal-level* cognitive processes.

In order for BRAINBOUND to pose a threat to the extended mind hypothesis, we need to read BRAINBOUND as a thesis about personal level cognitive processes, and in order for a thesis about personal level cognitive processes to pose a threat to the extended mind hypothesis, we need an argument to the effect that brain-bound, personal level cognitive processes have a special priority over extended, personal level cognitive processes. How might Adams and Aizawa argue for this? Their discussion of derived content suggests an answer. According to Adams and Aizawa, cognitive processes require non-derived representations, which they characterize as "representations that mean what they do independently of other representational or intentional capacities" (2010, p. 31). By contrast, "derived content arises from the way in which items are handled or treated by intentional agents" (p. 32). Insofar as numerals written on paper are representations with content, it would seem that the content of such representations is derived:

> Strings of symbols on the printed page mean what they do in virtue of conventional associations between them and words of language. Numerals of various sorts represent the numbers they do in virtue of social agreements and practices. The representational capacity of orthography is in this way derived from the representational capacities of cognitive agents. (2001, p. 48)

[11] See footnote 9.

The objection to calling calculations done on paper cognitive would then be this: insofar as the process involves states or items with content (e.g. numerals representing numbers), the content of those states or items is derived from the content of thoughts. Since a process is cognitive only if the process involves non-derived content, calculations done on paper are not cognitive, properly speaking.

Of course, the appeal to the non-derived/derived distinction works here only if Adams and Aizawa have the direction of derivation right, and they acknowledge that, "There are philosophers who think that meaningful language comes first, and then thoughts derive their semantic content from language" (p. 34). In the course of considering how this "language-first" proposal might apply to non-human animals, Adams and Aizawa offer a response to these philosophers:

> Perhaps vervet monkey calls have semantic meanings, such as that there is a predator above or a predator below, that do not derive their meaning from content-bearing mental states. Thereafter, vervet mental content derives from these calls. We resist this order of derivation, however, since the system of vervet monkey calls, like many systems of animal communication, does not appear to be sufficiently elaborate to do justice to the range and diversity of mental representations that vervet behavior suggests. There are few vervet calls, but many vervet thoughts. (2010, p. 34)

However, even if we grant that Adams and Aizawa are right about what vervet behavior suggests, it does not follow that the representations at issue here are representations attributable at the level of the whole vervet (the vervet analogue of the personal level). The argument for the existence of content not derived from language rests on the plausible claim that there is content relevant to vervet behavior that cannot be explained in terms of the content of vervet calls, but as far as this point goes, the additional content may be content involved in "sub-vervet" processes. The worry here is not that there is an in-principle objection to calling sub-personal content 'thought content'; rather, the worry is that in doing so, our theorizing will fail to honor important distinctions. Dennett is clear on this point:

> Should we call this internal information processing reasoning, or thinking, or are there some other phenomena that better fit our intuitions? If we prefer to heed the ordinary notion that reasoning is a matter of conscious acts of the mind, a better way to define reasoning would be as awareness1 of an argument sequence leading to a conclusion. The decision is parallel to the decision on whether 'aware1' or 'aware2' is the notion of awareness. Is introspective access or felicity of behaviour to be the benchmark of reasoning? Consider a mathematician who does a problem in his head without even saying the steps to himself, and when we ask him how he did it, he says 'I just knew'. Should we say he did the problem without thinking? He can tie his shoe without thinking, so why not solve the problem without thinking? Tying his shoe requires some information processing to go on, and so does solving the problem, and if we decide, implausibly, that this is what deserves the name thinking, then, of course, mute animals can think. If, on the other hand, we restrict thinking to something like 'consciously reasoning with concepts', then animals cannot think, since they cannot be aware1 of anything, but also people can do many quite intellectual things without thinking. (*C&C*, pp. 173–174)

Although Dennett would find it misguided to loosen the use of 'thought,' it can be harmless to do so. However, it is *not* harmless to use 'thought' in the expansive sense that includes sub-personal cognitive processes, go on to show that some

thought content is not derived from language, and then directly conclude that the contentful states and activities attributable to *people* are not derived from language.

By the lights of the explanatory project heralded by *C&C*, there is something fundamentally misguided about invoking this derived/non-derived distinction to argue for the priority of brain-bound, personal level processes over extended, personal level processes. Among the *explananda* of cognitive science are the intentional capacities of people, and the vision offered in *C&C* is that we can explain these capacities by analyzing them into sub-personal intentional capacities. The only way these analyses can accomplish their task is if we do not invoke the very contentful states and capacities we are trying to explain. The picture here is thus one where the intentional contents and processes of people depend on the existence of sub-personal contents and processes, so if Adams and Aizawa are correct that "Underived content arises from conditions that do not require the independent or prior existence of other content, representations, or intentional agents" (2010, p. 32), it will turn out that the intentional contents and processes of people – brain-bound *or* extended – are derived.

Of course, among the cognitive scientists and philosophers who want to defend the derived/non-derived distinction, perhaps many would accept that the locus of non-derived content is the sub-personal. As far as the debate over the extended mind hypothesis goes, however, such an admission would seem to undermine the dialectical force of invoking the distinction in the first place. The extended mind debate *seems* to be first and foremost a debate over whether the mind of a *person* extends beyond the boundaries of brain, and Adams and Aizawa *seem* to invoke the non-derived/derived distinction in order to show that the mind does not extend beyond the boundaries of the brain. However, if non-derived content is the mark of the cognitive, and it turns out that the intentional capacities of people depend on sub-personal intentional capacities, then *extended or not*, the intentional capacities of people are not cognitive. In this way, invoking the distinction may threaten to prove too much.

7.4 Conclusion

Dennett writes, "The problem of mind is not to be divorced from the problem of a person. Looking at the 'phenomena of mind' can only be looking at what a *person* does, feels, thinks, experiences" (*C&C*, p. 213). This is a reminder and a warning about how to do cognitive science, but it also invites a seemingly paradoxical conception of that science: the science of the mind is the science of the person, but many of the explanations that the science is poised to provide will not be couched in terms applicable to persons. The air of paradox is dispelled when the claim is properly clarified, of course, but like a Zen kōan, perhaps an important lesson lurks in this formulation. An account of the person will require a cognitive science that is rich with intentional characterizations of the brain, but you cannot locate the person

in the brain. Taking the extended mind hypothesis seriously requires taking the personal/sub-personal distinction seriously, and when we do, we discover that minds – and thus people – are large, and "the chessboard, the platform, the scholar's desk, the judge's bench, the lorry-driver's seat, the studio and the football field are among its places" (Ryle, p. 51).

References

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology, 14*(1), 43–64.

Adams, F., & Aizawa, K. (2010). *The bounds of cognition*. Malden: Wiley-Blackwell.

Clark, A. (2003). *Natural-born cyborgs*. New York: Oxford University Press.

Clark, A. (2011). *Supersizing the mind*. New York: Oxford University Press.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy, 72*, 741–765.

Cummins, R., & Roth, M. (2011). Intellectualism as cognitive science. In A. Newen, A. Bartels, & E. Jung (Eds.), *Knowledge and representation*. Stanford: CSLI.

Cummins, R., & Roth, M. (2012). Meaning and content in cognitive science. In R. Schantz (Ed.), *Prospects for meaning*. New York: de Gruyter.

Dennett, D.C. (1969/2010). *Content and consciousness*. New York: Routledge.

Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT Press.

Fodor, J. (1968a). *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House.

Fodor, J. (1968b). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy*, *65*, 627–640 [Reprinted in Representations (1981). Cambridge, MA: MIT Press].

Fodor, J. (1975). *The language of thought*. New York: Crowell.

Lycan, W. (1987). *Consciousness*. Cambridge, MA: MIT Press.

Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind*. New York: New York University Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion*. Pittsburgh: University of Pittsburgh Press.

Ryle, G. (1949). *The concept of mind*. Chicago: University of Chicago Press.

Chapter 8
Learning Our Way to Intelligence: Reflections on Dennett and Appropriateness

Ellen Fridland

Abstract In chapter three of *Content and Consciousness*, Dennett writes that the "capacity to learn from experience in such a way that […] behavior improves in prudence is what I shall call the intelligent storage of information". This statement amounts to a claim that learning functions as the criterion of intelligence, or, at least, the criterion for the intelligent storage of information. It is this connection between learning and intelligence that I defend in this essay. I begin by forwarding a definition of learning that combines a flexibility requirement with a success requirement. I then go on to argue that four features often cited as characteristic of intelligence: flexibility, transferability, manipulability, and appropriateness, are related to intelligence only insofar as they as they satisfy one of the two requirements of learning. Moreover, I argue that positing learning as the criterion of intelligence explains why there seems to be a natural connection between the above-listed features and intelligence. In the final section of the paper, I identify and categorize four different learning kinds. These categories correspond to distinctions that Dennett has proposed between Darwinian, Skinnerian, Popperian, and Gregorian creatures. Taken together, these considerations provide reason to accept that learning is the criterion of intelligence and that intelligence is a natural, biological, evolved phenomenon.

Unlike many topics introduced in Daniel Dennett's *Content and Consciousness*, the nature of intelligence has not become a central issue in the philosophy of mind.[1] To appreciate this fact, we should notice that the question about what makes a state, process or behavior intelligent is importantly distinct from another closely related

[1] To be fair, many philosophers such as Dretske (1988, 1990), Bermúdez (2003) and Hurley (2006) have explored importantly related issues. Yet, the majority of the work in this area of philosophy has been devoted to exploring the nature of representation, intentionality, propositional content, rationality, and information processing. A targeted conception of intelligence has not been offered as part of the philosophical literature.

E. Fridland (✉)
King's College London, London, UK
e-mail: ellenfridland@gmail.com; ellen.fridland@kcl.ac.uk

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_8

question about what makes a state, process or behavior psychological or mental.[2] The distinction between these two categories can be illustrated by way of noticing how philosophers and cognitive scientists use the labels "cognitive" differently. For example, for cognitive scientists, "cognitive" usually means something like "mental." Accordingly, perception, memory, learning, etc. are all necessarily cognitive phenomena. Philosophers, on the other hand, use "cognitive" to mean something like "intelligent" such that it makes sense to ask whether psychological phenomena like visual perception are cognitive or cognitively penetrable, as Fodor (1983), Pylyshyn (2000), Prinz (2006), and Siegel (2010) do. In short, some but not all mental phenomena are intelligent. Therefore, the question of what makes a phenomenon mental is different from the question of what makes that phenomenon intelligent. In this essay, I will pursue the question of what makes a state, process or behavior intelligent.[3] To do this, I will return to Dennett's initial proposal that learning and intelligence are intimately related phenomena.

8.1 The Goal

In chapter three of *Content and Consciousness,* Dennett writes that the "capacity to learn from experience in such a way that…behavior improves in prudence is what I shall call the intelligent storage of information."[4] This statement amounts to a claim that learning functions as the criterion of intelligence, or, at least, the criterion for the intelligent storage of information. It is this connection between learning and intelligence that I defend in this essay.

I begin by forwarding a definition of learning that combines a flexibility requirement with a success requirement. I then go on to argue that four features often cited as characteristic of intelligence: flexibility, transferability, manipulability, and appropriateness, are related to intelligence only insofar as they as they satisfy one of the two requirements of learning. Moreover, I argue that positing learning as the criterion of intelligence explains why there seems to be a natural connection between the above-listed features and intelligence.

---

[2] Accounts of information processing or symbol manipulation such as Newell and Simon (1976) and Stitch (1983) are examples of the latter.

[3] The notion of intelligence that I am pursuing is a scientific notion. As such, my methodology will not be conceptual analysis. In this kind of endeavor, if various counterintuitive consequences result from my account, these will not immediately count as a *reductio* of the position. After all, science is often counterintuitive. Still, I hope to illustrate that what we think of as intelligence is already, to a large extent, in line with the claims that I am making here. As such, I would like the notions of learning and intelligence that I put forward to correspond to ordinary intuitions as much as possible. However, I do not insist that if ordinary intuitions conflict with the account I am offering, then the account is wrong. On my approach, it may turn out that we have *empirical* or *methodological* reasons that trump our ordinary intuitions. Intuitions ought to be considered, but they ought not to be the final arbiters.

[4] Dennett (1969, p. 49–50).

In the final section of the paper, I identify and categorize four different learning kinds. These categories correspond to distinctions that Dennett has proposed between Darwinian, Skinnerian, Popperian, and Gregorian creatures.[5] Taken together, these considerations provide reason to accept that learning is the criterion of intelligence and that intelligence is a natural, biological, evolved phenomenon.

8.1.1 Learning: A Working Definition

I define learning as a process where, as a result of experience or reasoning, the behavior, mental processing, or representations of subjects change in some way that contributes to the satisfaction of their goal(s).

We should notice that the above definition has two requirements: a flexibility condition, which requires a change in behavior, representation or processing, and a success condition, which requires the change to contribute to the satisfaction of the agent's goal(s). These two conditions are each necessary and jointly sufficient for learning.

The above definition of learning is meant to be as broad and inclusive as possible, whilst remaining informative. Accordingly, my definition is both more demanding and more inclusive than the definition of learning commonly offered in psychology, where learning is defined as "a relatively permanent change in behavior due to experience."[6] First off, in contrast to the psychological definition, I remain neutral about whether learning occurs on the neuronal, cognitive, or behavioral level. This means that my definition can be accepted by psychologists, neuroscientists, and computer scientists alike. Secondly, by requiring learning to contribute to a goal, the definition I offer introduces a normative component to learning. This normative component allows us to distinguish learning from other kinds of relatively permanent changes that result from experience like PTSD or myopia.

Additionally, my definition has the virtue of leaving open a whole range of substantive questions, which ought not to be decided by fiat. For example, in order not to exclude anti-represenationalists, I stay neutral about how psychological states are realized.[7] For similar reasons, I leave the term "goal" unqualified. I take it that a goal may be realized in action or in thought; and it may be aimed at success or truth.

Also, I use the word "change" instead of "develop" or "improve" in order to avoid limiting learning to states and behaviors that produce an increase in the probability of goal satisfaction. I assume that some learning allows for lateral changes, perhaps increasing the ease of goal attainment or decreasing the energy expended in achieving a goal, without thereby making it more likely that the goal will be attained.[8] Lateral modifications that do not improve the chances of success, but do

[5] Dennett (1996a).

[6] *The Dictionary of Psychology*, 3rd ed. "learning."

[7] See Varela et al. (1991) and Noë (2004).

[8] See Millikan (2000) Chap. 4, for similar observations about ways of improving.

contribute to its ease or facility count as learning. Changes that make goal satisfaction less likely or more difficult are not learning. Just like one cannot learn that the earth is flat, because it is not, the development of a panic disorder is not a learned behavior, though, of course, it is often acquired through experience. It is precisely for this reason that learning remains a normative notion.

Lastly, I use the plural "subjects" rather than the singular "subject" in order to leave open the possibility of group or species learning.[9] It seems to me that determining the proper ontological limits for being a subject of learning should remain an open philosophical and empirical issue. As such, my definition of learning makes room for different possible interpretations of what it means to be a subject, or agent, of learning.[10]

8.2 The Features

It is my contention that the above definition of learning has the virtue of allowing us to see how features commonly associated with intelligence establish their relation to intelligence by satisfying one of the two requirements of the learning definition. That is, flexibility, transferability, and manipulability satisfy the flexibility condition while appropriateness satisfies the success condition. Moreover, satisfying either the flexibility or success condition alone is insufficient for guaranteeing intelligence. As such, in reviewing the features commonly associated with intelligence and examining their connection to intelligent behavior, processing and representation, we see that it is the contributions that these features make to learning that underpins their participation in and connection to intelligence.

8.2.1 Flexibility

In this section, I argue that though flexibility is relevant for ascriptions of intelligence, it is only relevant insofar as it underpins the changes that learning demands. That is, flexibility is connected to intelligence because flexibility satisfies the flexibility condition of learning, and learning is the criterion of intelligence. Moreover, my claim is that flexible states and behaviors alone, disconnected from the goals of

[9] See, for example, Gilbert (1989, 2004) on the plural subject, group minds, and group mental states and, e.g., Rupert's (2005) response.

[10] Some may have noticed that on the above definition, God turns out not to qualify as intelligent. After all, God knows everything and so he cannot learn anything new. God cannot change, since he's already perfect. Some may see this as a *reductio* of my position but I think the most appropriate response to this "problem" is to appeal to the familiar fact that one (wo)man's *modus ponens* is another's *modus tollens*. If it turns out that on the above definition God is not intelligent, then so much the worse for God.

a subject, do not ensure intelligence. However, once we put flexibility together with the satisfaction of a subject's goals, what we end up with is learning.

Flexibility often creeps into discussions of intelligence, cognition, and psychological explanation. In fact, it isn't uncommon for intelligent behavior to be contrasted with fixed, inflexible behaviors. As José Bermúdez writes in *Thinking Without Words*, "a distinguishing mark of the cognitive is that it is variant, and not stimulus–response."[11] He goes on to contrast fixed, rigid behaviors with cognitively integrated "behavior that is flexible and plastic and tends to be the result of complex interactions between internal states, learning and adaptations contributing and determining present responses."[12]

Bermúdez is not alone in citing flexibility as a defining feature of intelligence. Hurley (2006), when discussing animal cognition presents a compelling picture of animals as inhabiting islands of rationality. These islands exist only to the extent that there are degrees of freedom or flexibility on them. And when Clark and Karmiloff-Smith (1993) defend the necessity of representational change in the development of human cognition, they connect this requirement to the need for flexible and manipulable states at higher, more explicit, levels of representation. In short, connections between intelligence and flexibility arise regularly for different theorists with various objectives.

However, when we consider flexibility and its connection to intelligence, we should ask what it is about flexibility that makes it a feature of intelligent behaviors, representations and processes. If we take some time to consider it, it becomes clear that it is not flexibility itself that we value, but rather, that for which flexibility makes room.

This point is easy to demonstrate since flexibility alone does not even come close to guaranteeing intelligence. After all, we have absolutely no reason to think that a mere lack of rigid determination makes a behavior intelligent. In fact, many flexible behaviors, in this sense, prove to be profoundly stupid. Think of random behavior, the most flexible behavior one could find. Is there any reason to think that a random act will necessarily qualify as genuinely intelligent? Imagine driving to the grocery store and stopping in the middle of the road to dance the Mambo; of going into coffee shop and reciting The Emancipation Proclamation; of sending a package to a friend and including an image of a power drill, a description of a mountain range at sunset, and a spoon. The fact is that behaviors that are not called forth by the context, though flexible, are hardly paragons of intelligence. In fact, quite the opposite seems to be true: a behavior that is not connected to its context in some strong, systematic way is almost sure to be disqualified from the realm of intelligence.

At this point, we may be reminded of the paradox of free will. Where it would seem that determinism undermines freedom, as Hume convincingly argued, being uncaused or not determined in no way reestablishes it.[13] The same seems to go for intelligence—fixed or inflexible behavior seems to undermine intelligence, but

[11] Bermúdez (2003, p. 8).

[12] Bermúdez (2003, p. 9).

[13] Hume (1748, VIII), and Dennett (1996b, 2003).

random or disconnected-from-the-context behavior doesn't spawn intelligence either. We seek a certain kind of connection between environment and action for intelligent behavior. We need flexibility, but not unbridled flexibility. In short, intelligence requires appropriately constrained flexibility.

Notably, the requirement that intelligence is both flexible and appropriately constrained is equivalent to the claim that intelligence requires learning. In fact, appropriately constrained flexibility amounts to satisfying the two requirements set out in the above definition of learning: the flexibility condition and the success condition. And we can see why this is correct because, upon reflection, it becomes obvious that the value of flexibility is not just in giving us any old options, but in extending to us the possibility of selecting the best option given our goals and opportunities. That is, we don't just care about having options for the sake of having options; we care about how those options are related to achieving our goals. After all, if a creature could select between various alternatives, but selected in a way that was thoroughly disconnected from its ends and circumstances, we would deem it no more intelligent than if the creature had responded with one designated, rote, or fixed behavior. It isn't just pursuing different strategies that we care about; it is about having the freedom to pursue the best strategy. And this amounts to having the capacity to learn.

In short, we want intelligent creatures to adjust their strategies based on what will be in their best interest. It is the flexibility to change its course, to try and retry, to learn from experience, or improve based on its present position where intelligence arises. As such, it seems that the reason that we value flexibility as a property of intelligence is because learning requires a degree of flexibility in order to allow for the appropriate modification of states, processes, representations, and behaviors. And this means that it isn't flexibility by itself that we value, but rather, flexibility's role in making possible the changes that are requisite for learning. And since learning is the criterion of intelligence, we can see why it is that flexibility is often cited as a symptom or feature of intelligent processing and behavior. So, it turns out the flexibility is not itself the mark of intelligence but, rather, a necessary feature of learning, which is integrally tied to intelligence.

8.2.2 Transferability

Another feature that is frequently invoked as characteristic of intelligence is transferability or context generality.[14] Transferability can be though of as a particular kind of flexibility: a kind of flexibility that highlights our commitment to intelligent behaviors or states playing a general role in our cognitive economy. Transferability highlights that intelligent processes ought not be context bound or domain specific. Like flexibility, transferability will satisfy the flexibility condition of learning and, like flexibility, transferability alone will be insufficient to guarantee intelligence.

[14] See, for example, Hurley (2006) on the combination and recombination of means and ends, and Evans (1982).

To get a better grip on what transferability adds to our concept of intelligence, we can contrast transferability with flexibility. Whereas flexibility yields responses that can vary in a particular setting, transferable behaviors are those that can be applied and re-applied in various settings and circumstances. In short, we can think of flexibility as creating a space of options in a given context, whereas we can think of transferability as allowing those options or strategies to be applied in multiple contexts, modalities, and environments. Of course, we should note that transferability requires a degree of flexibility, since a fixed state or behavior could not break free from its role in one context in order to be transferred into others.

To see how transferability is related to intelligence, we can begin by looking a classic discussion of conceptual content. As Gareth Evans has famously argued, in order for an element of thought to qualify as a concept, it must be capable of playing multiple roles in various propositions. He writes,

> It is a feature of the thought-content *that John is happy* that to grasp it requires distinguishable skills. In particular, it requires possession of the concept happiness—knowledge of what it is for a person to be happy; and that is something not tied to this or that particular person's happiness. There simply could not be a person who could entertain the thought that John is happy and the thought Harry is friendly, but who could not entertain—who was conceptually debarred from entertaining—the thought that John is friendly or Harry is happy.[15]

One cannot have the concept of BLUE without being able to think of various blue things: a blue couch, a blue chair, and a blue sky. And one cannot have the concept SKY, if one isn't able to think of the sky as, e.g., blue, cloudless, infinite, etc. Being a concept requires the capacity to recombine. Another way of saying this is that paradigmatically intelligent states are not tied to one role or context but can be transferred or applied in multiple roles and contexts.

This kind of multiple role-playing seems naturally tied to intelligence since a state or behavior that is singular or narrow in the scope of its application doesn't intuitively strike us as very intelligent. For example, if I can add jellybeans but not matchsticks or sheep, then one would be right to doubt if I am really adding. Since adding is an operation that should not be limited to one sort of object or setting, whatever allows me to calculate the sum of jellybeans seems distinctly dissimilar from the cognitive processes involved in basic arithmetic.

Crucially, the emphasis on transferability points to the fact that we want intelligent states and behaviors to be widely available to cognition.[16] We insist that knowledge and skills are accessible to an agent in a large number of circumstances. But all of this simply seems to be a way of saying that transferability underwrites the capacity to appropriately apply what one knows or does in one situation to novel situations. And such wide applicability, context generality, or transferability, when combined with the need to contribute to the satisfaction of an agent's goals, is a straightforward appeal to learning: for requiring that we apply something that we know here, to change or improve the likelihood that we will attain some goal there.

[15] Evans (1982, p. 102–103).

[16] Of course, the exact degree of generality, wideness, or number of circumstances of application cannot be specified precisely.

After all, we should notice that, like flexibility, we value transferability for the sake of success or truth and not for itself. In the absence of enhancing or changing behaviors in one context by transferring knowledge and skills from another, that is, in the absence of learning or improvement, transferability seems quite useless. It would not do me any good to transfer what I have learned in yoga to map reading, unless it was going to contribute to the satisfaction of my map reading goals. Without a connection to my goals and the world, transferability would be as intelligent as random flexibility: which is to say, not very intelligent at all.

To end, it seems that transferability matters for intelligence because appropriately transferred behaviors and representations allow one to more easily reach one's goals. As such, we must admit that the ability to play multiple roles in multiple contexts isn't by itself a sign of intelligence, but only intelligent insofar as it is connected to the adaptability and modification of goal-directed behavior. In short, transferable behaviors satisfy the flexibility condition of our definition of learning, but in the absence of being appropriately tied to purposive behaviors, transferability falls short of ensuring intelligence. Importantly, because transferability does satisfy the first requirement of the learning definition, we can see why this feature is often taken to be characteristic of intelligence.

8.2.3 Manipulability

A third important characteristic that arises in philosophical discussions of intelligence is manipulability. We should notice that, like transferability, manipulability requires flexibility, since one cannot manipulate what one cannot change. And like transferability and flexibility, manipulability will be a particular way of satisfying the flexibility requirement of learning. All three features will also fail to yield intelligent behaviors in the absence of a condition tying them to the particular goals and context of the agent. As such, all three conditions must be combined with a success condition, and thus, to satisfy the definition of learning, if they are to guarantee intelligence.

Manipulability refers to the requirement that an agent herself, rather than the environment or some third party, is responsible for intelligent behavior and processing. "Manipulability highlights the fact that when we speak of intelligence we want behavior that is not only flexibly related to the world, but flexible as a result of its being under the control of an agent."[17] In this way, manipulabilty ensures that intelligent processes are top-down, hierarchical processes that an agent can plan, organize, reorganize, guide, and control.

Psychologists Richard Byrne and Anne Russon frame intelligence in terms of both flexibility and manipulability. They write,

> [W]e would be reluctant to describe as intelligent any sequence of behavior whose mental organization is a single unit or action connected to a goal representation, a long sequence of linear

[17] Fridland (2013, p. 212).

> associative connections or a rigid hierarchical structure. Thus whether a behavioral structure is modifiable by the individual becomes crucial in diagnosing it as "intelligent" (1998, p. 671).

And Prinz (2004) goes as far as to define cognition in terms of manipulability. He states, "[c]ognitive states and processes are those that exploit representations that are under the control of an organism rather than under the control of the environment."[18] For Prinz, organismic control, which in mammals involves the prefrontal cortex, is at the heart of intelligent processing.

One important implication that follows from the requirement that intelligent processes be manipulable is that intelligence becomes a personal-level phenomenon. This is because manipulability requires global, integrated, centralized, hierarchical processes that are not available to subpersonal systems. That is, to be manipulated, a state must be targeted by higher-order states or mechanisms. The requirement that intelligent states are personal-level accords nicely with our intuitions about intelligence since, at the very least, the requirement that behaviors, processes, or representations be manipulable puts intelligence in the same realm as, for example, rationality and knowledge.

At this point, however, we should ask whether being under the control of an agent is sufficient for intelligence. But again, as with flexibility and transferability, the answer must be "no." For similar reasons as those presented above, we see that simply being under the control of the agent, in the absence of a deep and systematic connection to the goals and environment of an organism, will not yield intelligence. That is, if manipulability is not going to contribute to the satisfaction of a creature's goals by selecting or choosing the appropriate strategies in diverse and dynamic circumstances, that is, if manipulability isn't going to foster learning, then it is not obvious why manipulability is relevant for discussions of intelligence.

After all, what good is top-down control, if it runs counter to or even just neutral with one's own interests? If I made various true assertions that were deeply disconnected from my setting and circumstances, would my control over these assertions be enough to make them intelligent? Would my statements be any more intelligent than a digital computer's central processor? The fact is that like flexibility and transferability, manipulable behaviors should not be *determined* by the environment, but they must be lawfully and meaningfully connected to it. Without this further condition, it is difficult to see why being under the control of the agent matters for being intelligent. Surely, if we see that the behaviors, representations or processes of a subject are consistently disconnected from the objectives and environment of the organism or system then their being manipulated by top-down processes is hardly sufficient for making them intelligent.

It seems that manipulability's role in intelligence is to ensure that learning, or the changes and improvements that allow a creature to satisfy its goals, are not simply the result of passive, externally determined responses. In this way, manipulability endows learning with an active, deliberate component. But it is learning that must have this active feature. That is, control alone without a connection to goals is not sufficient for intelligence.

[18] Prinz (2004, p. 45).

8.2.4 Appropriateness

What the above discussion makes clear is that in order to produce intelligent behaviors or processes, what needs to be added to flexibility, transferability, and manipulability is the appropriate grounding in an organism's needs and environment. As such, it may seem that it is appropriateness and not learning that constitutes the difference between an intelligent and unintelligent behavior. But as with the above features, appropriateness alone, that is, satisfaction of the success condition, without the capacity for change and improvement, that is, without the satisfaction of the flexibility condition, is insufficient to guarantee intelligence. An inflexible, nontransferable, or nonmanipulable behavior, though appropriate, is not sufficient for grounding attributions of intelligence. But this is simply to say that an appropriate behavior lacking the flexibility that when combined with it amounts to learning, is not intelligent.

In chapter three of *Content and Consciousness,* Dennett appeals to the notion of appropriateness in order to elucidate his claims about intelligence. He states that "[t]he criterion for intelligent storage is then the appropriateness of the resultant behavior to the system's needs given the stimulus conditions of the initial input and the environment in which the behavior occurs."[19] Dennett is right, of course, that appropriateness is central to intelligence, but it is important to clarify that it is only a flexible appropriateness that yields intelligence, proper.[20]

In line with Dennett's position, I suggest we understand "appropriateness" as a general term for getting something right, given one's goals and circumstances. Importantly, getting something right or doing the right thing can only be evaluated relative to a particular context. Saying, "Boston is the capital of Massachusetts," though true, isn't the right thing to say when the conversation is about cattle. And picking up a pen may be the right thing to do if one wants to write a check, but it is not the right thing to do if one is up to bat. It seems that no matter how clever or sophisticated a thought, action, or process is, without a connection to other states, behaviors or processes,[21] it simply cannot qualify as intelligent.[22]

As Dennett points out, "since appropriateness is not an intrinsic physical or formal characteristic of any thing or event, no examination of the relations between intrinsic characteristics of input and output will give us a clue about intelligence."[23] So, no behavior or representation, no matter how internally coherent or consistent

[19] Dennett (1969, p. 50).

[20] From the text, it is difficult to discern if Dennett takes his statement about appropriateness to qualify his previous assertion about learning, if he takes these two to be equivalent concepts, or if he takes appropriateness to be the more fundamental quality of intelligence.

[21] See Davidson (1975) for similar considerations about the relationship between language and thought.

[22] As Dennett has written, "The capacity to use and store information intelligently, then, does not emerge automatically at any degree of size or complexity of the information storage and processing mechanisms, but is an additional and separable capacity" (1969, p. 51).

[23] Dennett (1969, p. 50).

could qualify as intelligent, if that behavior does not bear the proper connections to other states and behaviors. As such, we should understand appropriateness as guaranteeing the following: that a behavior, representation or process is instantiated at the right time, place, and way given the goals of the creature and the affordances of its environment. And no behavior or state that doesn't have this feature qualifies as appropriate.

But is being appropriate sufficient for intelligence? I will argue that the answer to this question is "no." This is because, if a behavior cannot change appropriately in changing environmental conditions, that is, if a behavior is not capable of appropriate modification, then that behavior is not intelligent. I will argue for this claim in two moves: First, I will make clear that the notion of intelligence tacitly assumes appropriateness in contrary counterfactual circumstances, i.e., intelligence requires responding differently, if the situation were different. Second, the flux of the natural world guarantees that situations will be different. As such, in the natural world, intelligence requires the flexibility to change one's behavior appropriately. Put differently, intelligence requires the capacity to learn.

In order for a behavior, representation, or process to qualify as intelligent, it is not enough that it is instantiated at the right time, place and way, given the organism's needs and context. Though acting appropriately is an important feature of intelligence, I argue that there is an additional, tacit assumption involved in ascriptions of intelligence. This assumption can be formulated by appeal to Dretske's counterfactual condition for knowledge.[24] We can say that intelligent behavior requires that:

(CC) If *b* is not appropriate in context *c*, then *S* will not instantiate *b* in *c*.

The counterfactual condition rules out states that are only appropriate as a result of chance, luck, or accident from qualifying as genuinely intelligent.[25] Essentially, this condition affirms that intelligence requires a strong, systematic, and flexible connection between a behavior and its environment. This kind of connection can be established only if we incorporate a counterfactual condition because, sometimes, luck makes a behavior the right, appropriate, or successful behavior, even when it is not intelligent.

I'll elucidate this point with an example:

> A common piece of advice that college students pass along to their friends who stayed out partying instead of studying for their exams is to choose "c" for every answer on a multiple-choice test. The idea is that, at least some of the time, "c" will be the right, i.e., the appropriate, answer. But though this strategy may betray some intelligence (not a great deal, since studying would clearly be a more intelligent alternative) when the student chooses "c" as a response to a test question, she is not responding intelligently.[26] Not because "c" isn't the right answer (the point of the advice is to maximize the number of times that the student

[24] See Dretske (1969).

[25] One may argue that a state isn't appropriate if it doesn't meet CC. In this case, being appropriate would be equivalent to being flexibly appropriate. As such, the distinction between learning and appropriateness would vanish.

[26] This is why Dretske (1981) says that a broken clock is *not* right even once a day!

> will choose the right answer), but because the behavior cannot satisfy the counterfactual condition. That is, even if the right answer was *not* "c", the student would choose "c" anyway.

Intuitively, this helps us to see why ascriptions of intelligence require CC. We see that intelligence requires not just doing the right thing at the right time in the right place, given one's goals and needs, but also, not doing that thing if it is not the right time, place, way, etc. The reason why choosing "c" for every answer makes choosing "c," even when it is the right answer, not intelligent is because this behavior doesn't meet CC.[27] The behavior appears intelligent because it is appropriate, i.e., it is right, but on analysis, we conclude that it is not intelligent because it doesn't bear the proper systematic and flexible connections to the world. This is precisely the difference between the strategy of choosing "all cs" and the strategy of studying, learning the subject matter, and only choosing "c" when it is the right answer. The latter is intelligent while the former is not.

Once we have established that intelligence is not simply determined by appropriately responding to a situation, we can think about the kinds of demands that the natural world places on creatures. That is, we can think about what kinds of contexts a real creature will have to encounter and respond to appropriately. With only a moment's consideration, we should see that ecological contexts shift and change regularly. It is not simply that animals encounter bivalent scenarios: i.e., worm (w) or no worm (−w), but situations like (1/2w) where only part of the worm is visible, or (ww) where the worm is in water and not on land, or (mw) where the worm is in another bird's mouth. Each of these scenarios requires more than a simple, "on/off" mechanism in order for an animal to respond appropriately. Appropriateness in the natural world, as it turns out, requires a nuanced, flexible set of responses.[28]

---

[27] We can also think of Charlie Chaplin's *Modern Times* in this context. In particular, we can recall the scene when Chaplin goes from tightening the bolts on the conveyer belt, to using his wrenches to tighten anything they will fit, including the buttons on a lady's dress.

[28] A paradigm example of lacking this sort of flexibility is the wasp, *Sphex ichneumoneus:* "When the time comes for egg laying, the wasp *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into the burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralyzed cricket, which has not decayed, having been kept in the wasp equivalent of deep freeze. To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulnessDOUBLEHYPHENuntil more details are examined. For example, the Wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again she will move the cricket up to the threshold and re-enter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion this procedure was repeated forty times, always with the same result" (Woodridge 1963, p. 82). See also, Dennett (1996b).

As such, in order to respond appropriately to changing environmental conditions, that is, in order to respond appropriately in the natural world, a creature must be able to adjust its strategy based on its circumstances. And this is precisely what learning amounts to: it requires modifying or adjusting one's behaviors and representations in a way that will contribute to the satisfaction of one's goals. We see that without this kind of flexibility, success or appropriateness at a time does not get us very far in our quest for intelligence.

So, if a behavior only qualifies as appropriate in one context but is not sensitive or responsive to various relevant, graded, environmental changes, I think we'd be hard pressed to call that behavior intelligent. At the very least, that behavior would lack all of the features that we've cited above as characteristic of intelligence. But, as we saw above, those features without appropriateness don't get us very far either. However, if we take these features together, what we see is that they amount to learning. That is, they amount to the satisfaction of the flexibility condition and the appropriateness condition, which taken together constitute learning. So, if we take learning as foundational, we can see why appropriateness matters for intelligence, since no behavior, process, or representation could be an instance of learning if it were not appropriate but we can also see why flexibility, transferability and manipulability matter, too.

In light of the above, we see that the capacity to learn incorporates appropriateness with the three features of intelligence discussed above. Further, this criterion accounts for why these features seem to be characteristic of intelligence by highlighting their connection or contribution to learning. This means that the learning criterion both unifies and explains the features that we take to be characteristic of intelligence. Methodologically, it would seem that a substantive, unified, explanatorily powerful criterion of intelligence is exactly the one that we want.

8.3 The Learning Condition: Past and Future

Before ending, I'd like to be clear about how learning functions as the criterion of intelligence. My claim is that *either* past or future learning qualifies a behavior, process, or representation as intelligent. Therefore, if a state or behavior is the result of past learning or if that state or behavior serves as the basis for future learning, then the state or behavior shall qualify as intelligent. Satisfying either disjunct is sufficient for meeting the learning criterion. This means that learning as a criterion for intelligence is bidirectional or bi-temporal. This may seem like an odd qualification, but there are good reasons to think that it is required for an adequate account of intelligence.

First, we should note that past and future learning usually go hand in hand. That is, a behavior that is potentially modifiable by learning in the future is ordinarily a behavior that has been acquired through learning in the past. This fact seems to underlie Dennett's point that "more intelligent animals require longer periods of infancy, but gain in ability to cope with novel stimuli because of the proportion of

'soft-programming'—programming not initially wired in and hence more easily overruled by novel stimuli."[29] Essentially, we see that the capacity to deal with novel situations, that is, the capacity for learning, is often importantly related to a state's development through past learning. Past and future modifiability are both rooted in the potential for flexible, variable, and appropriate responses. It turns out that the opposite is also true: behaviors that are *not* acquired through experience or learning are often behaviors that do not have the potential to change as the result of experience and learning. Tropistic or reflexive behaviors are obvious examples of this kind of rigidity.[30]

Though the conjunction of past and future learning is often the norm, there are certain exceptions, for which it is important that we account. It is because of these exceptions that the learning condition should be formulated as a disjunction, rather than a commitment to either one of the disjuncts, or to their conjunction.

To start, there is the rather depressing reality that people peak, plateau, and die. For example, my gymnastics skills peaked during my sophomore year of high school—they've only gotten worse since then. And my math skills plateaued in college—in years, they have neither improved nor changed. And the inevitable is inevitable—nothing will change or improve after that.

These realities are important for us to consider since they highlight that future learning cannot be the sole criterion upon which we base ascriptions of intelligence. After all, we should not want a criterion that necessarily classifies "peak" or "near death" behaviors as unintelligent. But that is exactly what would happen if *future* learning (not just future *or* past learning) were necessary for intelligence.

In order to identify the cognitive nature of such events, we should have the opportunity to look backward to past learning. In this way, we can determine how sensitive and responsive these processes have been to experience, success, and failure. That is, we can assess whether the organism bears a non-arbitrary, systematic, meaningful connection to the world by specifying how its behaviors have been formed.

Just as future learning runs into hurdles as the sole criterion of intelligence, past learning faces challenges, too. For example, Prinz (2004) has argued against learning as the criterion for intelligence based on the presence of innate cognitive mechanisms. He states, "[i]t seems coherent to postulate innate cognitive abilities (cognitive scientists do that all the time), and innate abilities are, by definition, unlearned."[31]

In order to accommodate for intelligent mechanisms, abilities, or knowledge that are not the result of ontogenetic learning, I suggest we focus on whether such knowledge or abilities are subject to learning in the future. That is, we can ask if these processes have the disposition to change, improve, and develop over time and experience. In this way, using a counterfactual, we can evaluate them for their intelligence based on what kind of changes or improvements they make possible. Using

[29] Dennett (1969, p. 66).

[30] See Bermúdez (2003), and Dretske (1988) for more on these kinds of behaviors.

[31] Prinz (2004, p. 44)

this strategy, we avoid having to say that unlearned states are necessarily unintelligent,[32] and we get to hold onto the learning criterion, too.

The disjunctive learning condition also helps us to see that the reason that learning is tied to intelligence is not because we are particularly concerned with causal histories, but because causal histories tell us something important about the nature or constitution of the behaviors, representations, and processes that have them. The reason potential or future learning counts as a criterion of intelligence is because the disposition to learn tells us not only about the way that a state, process, or behavior is related to the world, but about the underlying qualities of that state that make it possible for it to be related to the world in that way. In short, having a bidirectional learning condition highlights both the extrinsic character of intelligence and the fact that having the right character is often connected to internal capacities, abilities, mechanisms, and systems.

8.4 An Objection and an Opportunity

The burning question at this point of the paper should be, of course: what kinds of changes qualify as learning? Does sensitization count as learning? How about habituation? All adaptions? Any useful modification at all? There are, after all, an enormous number of changes in behavior, processing and representation that contribute in some way to the satisfaction of a creature's goals. And many of these changes do not seem to be paradigmatically intelligent. This fact is the second reason that Prinz (2004) thinks that learning makes a poor criterion of intelligence. He states,

> …some insects are capable of learning and memory. Fruit flies, for example, can be conditioned to avoid electric shocks. We might even attribute learning and memory to individual neurons.[33]

To get clear on this issue, it may be helpful to return to an exchange between Dennett (1991) and Dretske (1990) where it is precisely the scope of learning about which they disagree. Dretkse insists that real learning is ontogenetic learning, that is, learning within a lifetime.[34] According to Dretske, only states that are the result of intra-generational learning really qualify as meaningful or intelligent. Dretske writes,

> In order to get meaning itself (and not just the structures that have meaning) to play an important role in the explanation of an individual's behavior (as beliefs and desires do) one has to look at the meaning that was instrumental in shaping the behavior that is being explained. This occurs only during individual learning.[35]

---

[32] Or, as will become clear below, only intelligent due to participating in a lower level of learning.

[33] Prinz (2004, p. 44).

[34] Natural selection gives us something quite different: reflex, instinct tropisms, fixed-action-patterns, other forms of involuntary behavior—behavior that is (typically) not explained in terms of the actor's beliefs and desires Dretske (1990, pp. 14–15). See also Dretske (1988, pp. 104–107).

[35] Dretske (1990, p. 14).

In contrast, Dennett asserts that that phylogentic learning is no less learning than learning within a lifetime; to cut it in any other way, he argues, is quite arbitrary. Dennett writes,

> The curious question, of how much traffic with the world is enough, somehow, to ensure that genuine learning has been established, is simply an enlargement of the curious question that has bedeviled some evolutionary theorists…But if nothing but an arbitrary answer (e.g., 42 generations of selection) could "settle" the question for natural selection, only arbitrary answers (e.g., 42 flies must buzz) could settle the question for a learning history, for the processes have the same structure. They must begin with a fortuitous or coincidental coupling, thereupon favored—and they have the same power to design structures in indirect response to meaning.[36]

It would seem that we are at an impasse. What appeared to be our best shot at a unified, explanatorily potent criterion of intelligence now seems too broad to adequately differentiate intelligent from unintelligent behaviors, representations and processes. It seems that either we have to allow the changes that result from natural selection, classical conditioning, habituation, sensitization, and everything in between, to qualify as learning, or we have to deny that learning alone can function as the criterion of intelligence.

In reality, things are not so bad. As opposed to giving up the learning criterion, I suggest that we take learning's ubiquity as an opportunity to connect higher-order, human-level intelligence with the rest of the natural world. Specifically, I suggest that we begin by differentiating various kinds of learning into clear and substantive categories. In doing so, we can offer non-arbitrary boundaries for different learning types and, thus, different levels of intelligence. Additionally, this approach will ground learning and intelligence in an evolutionary history.

All learning will turn out to be appropriate and flexible (to some degree), and at higher taxonomic levels, we'll begin to see transferability and manipulability emerge. We will not have to decide which learning level is "really" learning but by introducing substantive distinctions, we can produce clear boundaries between learning kinds. In this way, those theorists who want a more stringent criterion of learning can appeal to the learning level of their preference as the criterion of intelligence. Using this strategy, we can appease those theorists who want only higher-levels of learning to count as intelligent, without needing to abandon learning as our primary criterion of intelligence.

In order to categorize learning into different kinds, I suggest that we follow Dennett's (1996a) classification of creatures. That is, I suggest we categorize learning according to whether it is of the Darwinian, Skinnerian, Popperian or Gregorian variety. Crucially, distinguishing between these kinds of learning will allow us to differentiate between, phyologentic, ontogentic, representational, and self-conscious varieties of learning, making it possible to understand their respective connections and contributions to the evolution and development of intelligent systems.

[36] Dennett (1991, p. 125).

The category of *Darwinian learning* should include the systematic changes that take place via natural selection. This will be our lowest level of learning. Here we connect appropriateness or the success condition with only a small degree of flexibility. The flexibility of Darwinian learning is achieved via selectional processes over multiple generations and is measured in evolutionary time. A simple example of Darwinian learning is the camouflaging capacities of lizards, which have evolved to decrease the likelihood of predation.

Next, we have *Skinnerian learning*. Skinnerian learning is best understood as resulting from classical or operant conditioning. This kind of learning is trial and error and it comes down to a creature's capacity to "modify (or redesign) their behavior in appropriate directions as a result of a long, steady process of training or shaping by the environment."[37] As Dennett notes, "there is no doubt that most animals are capable of" this kind of learning. An example of Skinnerian learning would be developing a preference for red cups after having been given sugary drinks in red cups in the past. At the Skinnerian learning level, we have appropriateness and a bit of flexibility, but not all that much.

At the third level, we have *Popperian learning*. Popperian learning is learning that goes on inside the animal without necessarily first having gone on in the world. In contrast to Skinnerian learning, Popperian learning does not need to be acquired through a long process of actual reinforcement. Instead, such learning can result from weighing various options in one's mind, that is, it can result from doing trial and error in one's head. This is why this kind of learning is called Popperian—because it "permits our hypothesis to die in our stead."[38] Dennett thinks that "mammals, birds, reptiles, amphibians, fish and even invertebrates exhibit the capacity to use general information they obtain from their environment to presort their behavioral options before striking out."[39] We can think of the Popperian level of learning as exhibiting appropriateness, a medium to high degree of flexibility, some degree of transferability and, arguably, some degree manipulability as well. Ruth Millikan (2006) gives the example of a squirrel checking out a bird feeder from different angles, trying to figure out the best way up, as an example of Popperian learning.

The last and highest learning level is *Gregorian learning*. At this level we have creatures like us who can use language to reason, problem-solve, and learn.[40] This is the kind of learning that most of us do in school. It is at this level that we can naturally talk of explicit knowledge, agency, meta-representation and self-consciousness. At this level, we have appropriateness, and a high degree of flexibility, transferability, and manipulability. I should add that at this level we get creativity, too.

By categorizing learning kinds in the above fashion, we gain the capacity to distinguish between higher and lower forms of learning and, thus, non-arbitrarily decide where intelligence of the sort we care about comes into play. By using this approach, we need not give up learning as the criterion of intelligence since we can,

[37] Dennett (1996a, p. 87).

[38] Dennett, ibid., p. 88

[39] Dennett, ibid., p. 93

[40] Dennett, ibid., p. 99

if we wish to, identify only specific levels of learning as constituting our criterion of intelligence. However, we have the added benefit of naturalizing intelligence by connecting it to more basic forms of learning.

In this way, we won't have to decide who's right about learning: Dennett can have all the categories of learning, Dretske can take the Skinnerian variety on up, and Prinz can be at home with Popperian and Gregorian learning. If one prefers a higher level of learning as the "real" criterion of intelligence, then that is okay by me. It seems to me that these preferences don't add much to our understanding of the world as much as they betray what we want and like about it. But to get this far, I think, is to get to a good place. We have shown that intelligence is not simply in the eye of the beholder. We have laid out a clear, substantive criterion for intelligence, and also tied it to our evolutionary past. Not bad for a day's work.

References

Bermúdez, J. (2003). *Thinking without words*. Oxford: Oxford University Press.

Byrne, R., & Russon, A. (1998). Learning by imitation: A hierarchical approach. *Behavioral Brain Sciences, 21*(5), 667–721.

Clark, A., & Karmiloff-Smith, A. (1993). What's special about the development of the human mind/brain? *Mind & Language, 8*(4), 569–581.

Davidson, D. (1975). Thought and talk. In S. Guttenplan (Ed.), *Mind and language*. Oxford: Oxford University Press.

Dennett, D. C. (1969). *Content and consciousness* (2nd ed.). London: Routledge & Kegan Paul.

Dennett, D. C. (1991). Ways of establishing harmony. In B. P. McLaughlin (Ed.), *Dretske and his critics*. Cambridge, MA: Blackwell Publishing.

Dennett, D. C. (1996a). *Kinds of minds*. New York: Basic Books.

Dennett, D. C. (1996b). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Dennett, D. C. (2003). *Freedom evolves*. New York: Penguin.

Dretske, F. (1969). *Seeing and knowing*. Chicago: University of Chicago Press.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.

Dretske, F. (1990). Does meaning matter? Postscript. In E. Villanueva (Ed.), *Information, semantics, and epistemology*. Cambridge, MA: Blackwell.

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fridland, E. (2013). Imitation, skill learning, and conceptual thought: an embodied, developmental approach. In L. Swan (Ed.), *The origins of mind: book Series in Biosemantics*. Springer: Dordrecht.

Gilbert, M. (1989). *On social facts*. London: Routledge.

Gilbert, M. (2004). Collective epistemology. *Episteme, 1*, 95–107.

Hume, D. (1748). *An enquiry concerning human understanding,* reprinted in 1999. Oxford: Oxford University Press.

Hurley, S. (2006). Making sense of animals. In S. Hurley & M. Nudds (Eds.), *Rational animals?* Oxford: Oxford University Press.

Millikan, R. G. (2000). *On clear and confused ideas*. Cambridge: Cambridge University Press.

Millikan, R. G. (2006). Styles of rationality. In S. Hurley & M. Nudds (Eds.), *Rational animals?* Oxford: Oxford University Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery, 19*, 113–126.
Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.
Prinz, J. (2006). Beyond appearances : The content of sensation and perception. In Tamar Gendler & John Hawthorne (Eds.), *Perceptual experience* (pp. 434–460). New York: Oxford University Press.
Pylyshyn, Z. W. (2000). Is vision continuous with cognition? *Behavioral and Brain Sciences, 22*(3), 341–365.
Rupert, R. (2005). Minding one's cognitive systems: When does a group of minds constitute a single cognitive unit? *Episteme: A Journal of Social Epistemology, 1*(February), 177–188.
Siegel, S. (2010). *The contents of visual experience*. Oxford: Oxford University Press.
Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge, MA: MIT Press.
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
Woodridge, D. (1963). *The machinery of the brain*. New York: McGraw Hill.

Chapter 9
The Intentional Stance and Cultural Learning: A Developmental Feedback Loop

John Michael

Abstract In this paper, I propose a developmental explanation of the reliability of the intentional stance as an interpretive strategy, and by doing so counter an objection to Dennett's intentional stance theory (i.e. the 'If it isn't true, why does it work?' objection). Specifically, young children's use of the intentional stance enables them to learn from and thereby to become more similar to the adults in their culture. As a result, they themselves become increasingly intelligible to other people taking the intentional stance. Thus, the intentional stance and cultural learning constitute a feedback loop that (partially) explains the reliability of the intentional stance, and does so – contra Dennett's realist critics – without appealing to a realist interpretation of the descriptions speakers attach to intentional terms. However, I also suggest that this developmental perspective provides grist to the mill for a *causal* realist interpretation of the reference of intentional terms, insofar the causal interaction between intentional interpretations of behavior and cognitive development provides an anchor that links intentional terms to functional and/or neural processes. Importantly, causal (as opposed to descriptive) theories of reference make it possible to argue that intentional discourse can be referentially anchored to the causal machinery that produces behavior without generating true descriptions of it. I conclude by drawing out some consequences of the developmental perspective for the way in which we conceptualize the assumption of rationality that is at the core of the intentional stance theory.

9.1 Introduction

In this paper, I propose a developmental explanation of the reliability of the intentional stance as an interpretive strategy, and by doing so I counter an objection to Dennett's intentional stance theory. I will begin (Sect. 9.2) by briefly recapitulating Dennett's theory, as well as one of the central objections that has animated critical discussions of it,

J. Michael (✉)
Central European University, Budapest, Hungary
e-mail: johnmichaelaarhus@gmail.com; michaelJ@ceu.hu

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_9

namely that its instrumentalist character prevents it from accounting for the reliability (perhaps even indispensability) of the intentional stance as an interpretive strategy (Richardson 1980; Bechtel 1985; Fodor 1985; Millikan 1993; Fodor 1985 refers to this as "the 'If it isn't true, why does it work?' problem"). The bulk of the paper will then be devoted to articulating my own proposal.

The core of my proposal is the idea that young children's use of the intentional stance during cognitive development enables them to learn from and thereby become more similar to the adults in their culture, whereby they themselves become increasingly predictable and intelligible for other people taking the intentional stance. Thus, the intentional stance and cultural learning constitute a feedback loop that (partially) explains the reliability of the intentional stance. This proposal commits me to the following three claims, which I will present and defend in turn:

(i) Children take the intentional stance from early infancy (Sect. 9.3);
(ii) Doing so enables cultural learning (Sect. 9.4);
(iii) Cultural learning (partially) explains the reliability of the intentional stance (Sect. 9.5).

After presenting and defending these claims, I will then (Sect. 9.6) consider the implications of this developmental perspective for the reference of intentional terms. It will become apparent that the proposed feedback loop helps to meet the "If it isn't true, why does it work?" objection and thus to defend Dennett against his realist critics, i.e. it helps to explain why the intentional stance is an effective interpretive strategy without maintaining that adequate intentional explanations of behavior must be true statements about the causal machinery that produces behavior.

It is important to state at the outset that this is not intended as a *general* defense of Dennett's theory or as an argument against realism; it simply rebuts an objection that would otherwise threaten to undermine Dennett's theory. Thus, the account is compatible with various realist positions. In fact, I will be suggesting that the account opens the door to a *causal* realist interpretation of the reference of intentional terms, since the causal interaction between intentional interpretations of behavior and cognitive development provides an anchor that links intentional terms to functional and/or neural processes. Importantly, causal (as opposed to descriptive) theories of reference make it possible to argue that intentional discourse can be referentially anchored to the causal machinery that produces behavior without generating true descriptions of it. Thus, a causal realist interpretation would enable one to resist the strong realist inference that Dennett wants to resist while doing justice to the realist intuition that intentional discourse must be somehow anchored to the functional and/or neural processes underlying behavior.

I will close (Sect. 9.7) with some reflections on how the developmental account being offered here relates to Dennett's own evolutionary response to the "If it isn't true, why does it work?" objection, and to how it bears upon the assumption of rationality at the core of the intentional stance.

9.2 The Intentional Stance

The core of Dennett's intentional stance theory is the proposal that our everyday practices of explaining and predicting other people's behavior are best characterized in terms of what he calls the *intentional stance*. Taking the intentional stance toward a system (such as another person) is to approach it as an entity "whose behavior can be predicted by the method of attributing beliefs, desires, and rational acumen" (1987: 49). More specifically, an interpreter takes the intentional stance not just by ascribing any old beliefs and desires but by assuming that the system has the beliefs and desires it *ought* to have, and that it reasons rationally from *these* beliefs and desires (Dennett 1987, 2008).[1] In identifying the appropriate beliefs and desires, and working out their effects upon behavior, Dennett proposes that interpreters are guided by the following "rough-and-ready" principles:

(i) A system will have the beliefs it ought to have, given its perceptual capacities, epistemic needs and biography;
(ii) A system will have the desires it ought to have, given its biological needs and the most practicable means of satisfying them;
(iii) A system will perform the actions that it would be rational to perform, given its beliefs and desires (1987: 49).

There are a few key features of the intentional stance that are worth emphasizing. First of all, it is *normative* insofar as it accords a central role to a regulative ideal of rationality, in light of which interpreters try to make sense of target agents' behavior, i.e. they aim to construe it as maximally rational. Secondly, it is *holistic*: the various ascriptions mutually constrain each other in order to retain overall consistency. Thirdly, it is an *idealizing* method, since it assumes rationality as a regulative ideal, although we are of course not really perfectly rational creatures. Rather, we approximate that ideal well enough for the method to be useful (more on this in a moment).

For these reasons,[2] Dennett does not think that interpreting a target agent's behavior on the basis of the intentional stance entails postulating causally salient structures inside the head. In other words, interpretations do not generally aim to pick out the internal physical workings of the system that causally bring about the behavior: "…the beliefs and desires that it (folk psychology) attributes are not – or need not be – presumed to be intervening states of an internal behavior-causing

[1] Dennett distinguishes various stances that one can take in interpreting the behavior of different systems (Dennett 1971, 1978: 237–238, 1987). Apart from the intentional stance, one can, for example, take the *physical stance*, and assume that a system will be predictable on the basis of physical laws. Or one can take the *design stance*, and treat the system as the product of evolutionary or human design, in which case one will interpret the workings of its parts in terms of their likely functions.

[2] There are also some other reasons, to discuss which would be beyond the scope of this paper. See, for example, Bechtel (1985) for a discussion of them.

system" (1987: 52). Rather, for an agent to have a particular belief is merely for the attribution of this belief to be compelling to an interpreter, where an interpreter has a characteristic viewpoint and set of goals.

Does this mean that intentional ascriptions are purely relative, that intentional terms do not refer to anything objective, and/or that beliefs and desires are not real? Dennett has consistently resisted such radical consequences, and has therefore tried over the years to carve out an ontological middle ground between realism and eliminativism (sometimes calling it "instrumentalism"[3] and sometimes "mild realism"[4]). In fleshing out the sense in which his position is instrumentalist, Dennett has drawn upon the distinction between *abstracta* ("calculation-bound entities or logical constructs") and *illata* ("posited theoretical entities") (1987: 53), and suggested that beliefs and desires are best understood as *abstracta*. Thus, they are not like dark matter, the existence of which we take to be probable but uncertain because of various observations and theoretical considerations. Rather, they are more like centers of gravity, the existence of which is not a matter of probability but of convention. The question arises, then, if intentional states are abstract objects, just what they are abstractions *from*. The two obvious candidates are brain states and processes, on the one hand, and behavior on the other. Dennett himself favors the latter. In fact, he explicitly contrasts his view with the former view, which he attributes to Fodor, namely that intentional terms pick out "a pattern of structures in the brain," (ibid.: 191). On Dennett's view, intentional concepts pick out "real patterns" that are "discernible in agents' (observable) behavior" (2008: 191). The proof of their reality is that recognizing them makes it possible to formulate generalizations and predictions that one could not otherwise formulate.

Dennett has been careful to emphasize, however, that he is not an instrumentalist about psychological or neural states *in general*. Thus, he speaks of his "realism about brains and their various neurophysiological parts, states, and processes" (Dennett 1987: 72), and confirms his agreement that "it is reasonable to consider sensory experiences to be real states of the brain, states whose neurobiological properties will be discovered as cognitive neuroscience proceeds" (Dennett 1993: 210). In fact, he describes himself as being "as staunch a realist as anyone about those core information-storing elements of the brain, whatever they turn out to be, to which our intentional interpretations are anchored" (1987: 70). He just does not expect that those elements will turn out to be "recognizable as the beliefs we purport to distinguish in folk psychology" (1987: 71).

It seems fair to ask, then, just how our intentional concepts are "anchored to" those information-storing elements of the brain that cause behavior if intentional states are abstractions from behavior. And, indeed, realists have queried whether it is possible to account for the reliability (perhaps even indispensability) of the inten-

[3] E.g. in "Three Kinds of Intentional Psychology", Dennett (1987: 53).

[4] E.g. in "True Believers", Dennett (1987: 28), and in "Instrumentalism Reconsidered" Dennett (1987: 71).

tional stance as an interpretive strategy without endorsing a more robust realism about the intentional states and rational thought processes that it postulates. In other words, the empirical success of predictions based upon intentional state ascriptions is mysterious if intentional states do not really cause behavior. Fodor (1985) refers to this as "the 'If it isn't true, why does it work?' problem" (Cf. also Bechtel 1985; Millikan 1993; Dretske 1988).

Although Dennett resists the realist argument that recognizing patterns in behavior would only be useful if they corresponded isomorphically to a second set of patterns within the brain (2008: 201), his talk of "anchoring" does appear to acknowledge that, in order to be useful in explaining and predicting behavior, intentional terms must *somehow* be related to the brain states that cause behavior. In spelling out this relationship, Dennett appeals to evolution, arguing that *evolution is likely to have shaped us in such a way that we approximate the rational agents that the intentional stance posits.* Note that this proposal does not entail a commitment to an isomorphic relation between the concepts and distinctions that structure intentional discourse and the functional or neural levels of description. The idea is that evolution will have selected for *any* functional and/or neural mechanisms that lead to approximately rational behavior, and since there are in principle lots of different mechanisms that could achieve this, which ones actually underlie a particular pattern of behavior will depend on the details of evolutionary history. Thus, it would be foolish to expect to be able to extrapolate this in a straightforward fashion from descriptions couched in intentional terms.

However, although various functional and/or neural processes could underlie a pattern of behavior that is nevertheless one and the same pattern as described in intentional terms, one may be inclined to think that those functional and/or neural processes must produce some of the same effects in order to instantiate the same pattern. Take, for example, the following pattern: Jim *sees* an object O being placed at location L, he *hears* it being placed there, he is *told* that it is there, he acquires good reasons to *infer* that it is there, etc. We see a pattern in these cases, namely that Jim acquires the belief that O is at L. Surely, the neural processes giving rise to this belief are different in these different cases, and some of the functional properties will also be different. *But some of the functional properties will also be the same*: Jim will be led to infer that O is not at some other location, to desire to go to L if he desires O, etc. So it seems reasonable to expect that in Jim's brain, the information he attains in these different cases is going to be treated as equivalent regardless of its source, and that this is why it is fruitful to treat these cases as constituting a pattern.

This, at any rate, is the realist intuition. I will be attempting later on show that there is a way of doing more justice to this intuition than Dennett has so far done while stopping short of the standard realist inference that adequate intentional explanations need to be isomorphic with adequate functional or neural explanations. For now, however, let us round out the discussion of the intentional stance theory by briefly considering three sources of concern other than the "If it isn't true, why does it work?" objection:

1. *What about aberrant beliefs and desires?* The first concern is that there is some question as to how central the assumption of rationality is in our interpretations of others' behavior, given that we routinely ascribe beliefs and desires to people that depart from ideal rationality and therefore must appear aberrant from the perspective of the intentional stance theory. We might distinguish three categories of such beliefs and desires. First, there are beliefs and desires that agents ought not have given their perceptual access, epistemic needs and biography, or their biological needs (cf. the three rough-and-ready principles referred to above as constituting the core of the intentional stance). Secondly, there are beliefs and desires that are formed through faulty inferences. Thirdly, there are beliefs and desires that bear utterly non-rational connections to behavior. An example from Stich (1981) illustrates the first category: we may ascribe to Sam the desire for a chocolate bar even if we have already ascribed to him the desire to stay healthy and the belief that he (himself) has a nasty allergy to chocolate. Although it is not irrational on the part of Sam to have this desire (in particular if he actively resists the urge to act upon it), it is a desire that he ought not have given his biological needs, and is therefore in conflict Dennett's second "rough-and-ready principle" (see above). One might therefore say, tongue-in-cheek, that Mother Nature was irrational in endowing him with this desire. Secondly, even when they start out from perfectly good beliefs and desires, people sometimes make less-than-ideal inferences. For example, most people have the intuition that a politically active young woman with a college degree and feminist political views is more likely to be a feminist and a bank teller than just a bank teller (Nichols and Stich 2003: 145–147; Tversky and Kahneman 1983; Cf. Nisbett and Ross 1980). Moreover, when people are guided by their emotions, they tend to make all manner of dubious inferences (think of Othello). The problem, of course, is that such failures of rationality are common enough to be predictable and even understandable. Thirdly, we also routinely generate perfectly good explanations of others' behavior by ascribing beliefs, desires and other intentional states on the basis of their utterly *non-rational* effects upon behavior (cf. Goldman 2006, Chap. 3). For example, the poker player's twitch reveals that she is bluffing, the young man's blushing reveals that he is embarrassed or in love, etc.
2. *Can evolution underwrite an assumption of ideal rationality?* A second issue – also raised by Stich (1981) – is that then even if the assumption of ideal rationality were empirically adequate, evolution may not explain that adequacy. For evolution does not select for true beliefs but merely useful ones. Sometimes – perhaps often – false beliefs may be more useful than true/rational ones. To take an example from Stich (1981): if organisms in an environment with an abundance of food discover that some particular yellow fruit is poisonous, they may do quite well with a strategy of assuming all yellow fruit to be poisonous and avoiding it ("better safe that sorry"). Whatever one thinks of this example, it does seem plausible that evolution should favor some useful but false beliefs (and some useful but rationally sub-optimal belief-forming processes).

3. *Rationality underdetermines ascriptions.* A third source of concern is that the assumption of rationality at the core of the intentional stance theory – irrespective of its accuracy and its justification – underdetermines most everyday intentional ascriptions. Consider neutral beliefs and desires: Apart from beliefs and desires that one ought not have, we also routinely ascribe beliefs and desires that are *rationally neutral* – that is, there is no reason why one ought to have them. We ascribe to young children, for example, the belief that the Easter Bunny has hidden some eggs in the garden. Similarly, we routinely ascribe neutral desires to people, such as the desire to watch television, or the desire to drink some coffee. Moreover, there is always the problem of figuring out which beliefs and desires people ought to have, given that they desire what it makes evolutionary sense to desire. For example, in one particular culture, you should desire shells, because you can use shells to acquire food and other useful things, whereas in another culture you should desire money, because money is what is used to acquire food, etc. Thus, cultural knowledge is required in order to reach a level of specification that is useful for predicting/explaining/influencing people's behavior in everyday life.

In fairness, it must be noted that Dennett has in fact always acknowledged these kinds of limitations, and avowed that the intentional stance is supplemented and sometimes corrected with some empirical generalizations that people learn inductively (1987: 54). After all, he does not claim that the intentional stance theory captures all of folk psychology, but only the core, rationalizing part of it. Thus, it is really no problem for him to accept that people also sometimes explain and predict behavior by ascribing aberrant beliefs and desires[5] or to non-rational links between intentional states and behavior. The under-determination objection is perhaps trickier, insofar as it seems to imply that explanations and predictions of others' behavior *always* depend (in part) upon cultural knowledge that has nothing to do with rationality. Thus, cultural knowledge is not just an additional interpretive tool that complements the intentional stance but is required in order to apply the intentional stance.

None of these critical observations constitutes a knock-down objection, nor indeed does the "If it isn't true, why does it work?" objection. In raising them, my intention is to set out some of the unresolved issues that it would be desirable for an account such as mine to resolve. And I will be trying to show later on (Sects. 9.6 and 9.7) that the developmental perspective outlined here (Sects. 9.3, 9.4, and 9.5) generates novel and satisfying responses to these objections in a way that is largely compatible with Dennett's theory.

[5] This applies not only to beliefs and desires that are formed through faulty inferences but also beliefs and desires that agents ought not have, given their perceptual access, epistemic needs, biography and biological needs, and to beliefs and desires that have non-rational connections to behavior.

9.3 The Intentional Stance in Early Infancy

The first step in my argument will be to review some evidence that children take the intentional stance from early infancy. One potential obstacle to establishing that infants take the intentional stance sufficiently early for it to enable cultural learning is that children do not generally pass explicit verbal false belief tests until they are over 4 years of age (Wimmer and Perner 1983; Griffin and Baron-Cohen 2002; Apperly 2011) – if children do not pass this litmus test for theory of mind, or mind-reading, it may seem unlikely that they ascribe beliefs, desires and rational thought processes in early childhood.

But taking the intentional stance does not require infants to *conceptualize* or to *explicitly* ascribe mental states. Rather, it only requires that the expectations they form about other agents' behavior reflect a sensitivity to those agents' goals, their strategies for attaining those goals, and/or basic mental states such as attentional states and emotions. And there is a wealth of research in developmental psychology suggesting that this is the case. By 6 months, infants' gaze following reveals a sensitivity to attentional states (Senju and Csibra 2008). In fact, Reddy has argued persuasively that 2-month-olds respond to others' attention in ways that suggest that they experience others as attentional beings (e.g. Reddy 2003). By around 6 months, the phenomenon of affect attunement attests to a sensitivity to others' emotions (Stern 1985/1998); by 6.5 months infants perceive goal-related movements on the part of geometric shapes (Gergely and Csibra 2003); by 9 months, they distinguish cases where an agent does not do something because she is not trying from cases when she is trying but unable (Behne et al. 2005); by 10 months, they parse streams of behavior into units that correspond to what adults would see as separate actions (Baldwin et al. 2001).

Moreover, although the claim that infants take the intentional stance does not imply that they use the concept of belief or other mental states to understand others' behavior, some theorists argue that there is such evidence (Baillargeon et al. 2010; Carey 2009), and I think that a strong case can be made that they at least *partially* master such concepts by the end of the first year, and that the case begins to get quite strong by around 18 months. At 9 months, for example, children already grasp something about the relations among beliefs and desires, as evinced by their expectation that agents will be happy when a goal is achieved and disappointed when the goal is not reached (Tomasello et al. 2005: 6). In a study involving 15-month-olds, Träuble et al. (2010) used an apparatus designed such that an agent could cause a ball to be transferred from one bucket to another by manipulating the apparatus without seeing it (i.e. with her back turned). The finding was that infants expect an agent not to have a false belief even though she did not see the object transfer because she was turned the other way. This demonstrates an impressive ability to reason flexibly about the effects that various kinds of evidence (even non-perceptual evidence) will have on agents' beliefs. In other words, the infants must recognize a pattern insofar as they must interpret the adult agent's manipulation of the apparatus as being relevantly similar to (and thus constituting a pattern with) the agent's visual

perception of the location of the ball. Similarly, Song et al. (2008) found that 18-month-old infants' expectations are modulated if the experimenter *communicates* to the agent that the ball has been moved but not if she says merely that she likes the ball.

As already noted, taking the intentional stance does not presuppose this level of sophistication. In fact, Zawidzki reserves the term "intentional stance" for non-mentalistic interpretation. In characterizing the intentional stance, he writes: "Such a framework is not meant as a model of psychological processes; it is a framework for interpreting *behavior*, not mindreading" (Zawidzki 2013: 38). Although I think it is far from clear that Dennett really wants to restrict the term in this way, I will not dispute this, since Dennett exegesis is beside the point here, and Zawidzki is free to restrict the term in this way irrespective of whether Dennett would approve. But I will not adopt the proposed restriction: it seems to me that the intentional stance theory can be interpreted as an *analysis* of sophisticated mental concepts rather than an alternative to them, and that is the interpretation I will be working with here.

9.4 The Intentional Stance and Cultural Learning

The next step is to establish that taking the intentional stance enables cultural learning. Tomasello et al. (1993) distinguish three types of cultural learning that are either unique to humans or at least far more pronounced in humans than in any other animals, and which, crucially, depend upon learners and teachers understanding each other as beings who 'have intentional and mental lives like their own' (Tomasello 1999; 7): imitative learning, collaborative learning, and instructed learning.

Consider imitation. In Tomasello's (Tomasello 1999; Tomasello et al. 2005) terminology, what distinguishes imitation from emulation is that the learner focuses not only on the environmental effect of an observed action but on the observed agent's goals and strategies. This allows the learner to understand the agent as rationally selecting an appropriate sequence of actions to realize a goal. From about 18 months, infants tend to imitate incomplete but intended actions rather than replicating the exact behavior they have seen, e.g. when an agent tries but fails to close a drawer (Meltzoff 1995). Also, around 14 months, they selectively imitate features of an action that are relevant to the goal of the action – unless the manner in which the agent performed the action appears irrational given the goal, in which case they imitate the particular manner. Thus, for example, if an agent uses her head to turn on a light because her hands are occupied, the child will use his hands to turn on the light, presumably understanding that the adult rationally chose to use her head only because her hands were unavailable. If, however, the adult uses her head even though her hands are free, the child will tend to use his head too (Gergely et al. 2002). The interpretation offered by the authors of this study is that in the first condition (with the hands occupied), the child thinks that the adult only used her head

because her hands were occupied, i.e. that she would otherwise have used her hands, given that using the hands would have been the most efficient strategy. In the latter case, however (with the hands free), the child can discern no such reason why the adult used her head, and he therefore reverts to a default assumption that the adult is teaching him something new, i.e. something that he does not understand yet. In sum, the authors maintain that the children in the study were interpreting the agent's behavior in terms of goals and rational strategies to attain those goals, and that their imitative learning was guided by this interpretation.

Apart from acquainting children with new activities and objects that are common in their culture, imitation also shapes their development in numerous subtle ways. One aspect of moral development, for example, is the acquisition of appropriate behaviors for consoling others in distress, and there is evidence that young children tend to imitate the consolatory behaviors that have brought relief to them in the past when confronted with other individuals in distress, such as offering soothing physical contact or presenting objects that provide comfort or distraction (Hoffman 2000). It is true that this sort of imitation may not always require an understanding of the intention to console or of the emotional state of the person to be consoled. However, it is telling that around 18 months, when most children are able to make a distinction between self and other, as evinced by their ability to recognize themselves in a mirror (Lewis and Brooks-Gunn 1979), they also begin to react with empathic and sympathetic responses to victims of distress, and with appropriate, other-directed comforting and prosocial behavior (Bischof-Köhler 1991; Zahn-Waxler et al. 1992; Eisenberg et al. 2006; Vaish et al. 2009). This strongly suggests that infants in the second year of life not only react to others' emotions, as might occur in the case of emotional contagion, but do so in a way that reflects a sensitivity to the fact that some specific other person is experiencing some specific kind of emotion. More generally, an understanding of moral norms depends upon detection of others' mental states (especially emotions) in order to recognize when an action has caused distress to others, when adults have disapproved of actions, and when another individual is experiencing a negative emotion and is thus a candidate for consolation (Cf. Prinz 2005).

To turn to a different kind of cultural learning, namely instructional learning, there is also evidence that the intentional stance plays a crucial role here. In particular Csibra and Gergely (2011) argue that a "teleological" stance[6] enables children to interpret pedagogical cues from adults as indicating an intention to instruct them.[7] Thus, eye contact and other ostensive signals cause young children to expect to be imparted with generic shared knowledge and to adopt appropriate learning strategies. For example, eye contact leads infants to pay preferential attention to generalizable kind-relevant features of objects that adults refer to (Futó et al. 2010; Yoon et al. 2008), and to learn actions that incorporate causally opaque means (Gergely et al. 2002).

[6] Which they explicitly associate with Dennett's intentional stance (Gergely and Csibra 2003).

[7] The flip-side of this is that adults must take the intentional stance toward infants in order to treat them as candidates for learning. I will return to this point later (see Mameli 2001 and McGeer 2007 for thorough discussions).

One striking demonstration of the effects of ostensive signals on learning strategies was provided by Topal and colleagues (2008), who propose that children's perseverative search errors in the A-not-B task may be due to a response on their part to ostensive signals made by the adult experimenter – specifically, eye contact may elicit a "pedagogical learning stance" (Topal et al. 2008: 1832), which leads the infants to expect that the adult intends to teach them some generalizable information, such as that the object is generally located in location A, or that one generally searches for the object in location A. This heuristic then distracts the children from making use of the evidence they have just seen that the object has been moved to location B, and thus leads them to make the perseverative error. And indeed, in a version of the task that does not involve the experimenter attending directly to the child, the perseverative search errors were dramatically reduced.

Understanding others as intentional agents with goals and strategies is also necessary for children in learning how to use tools and symbols, since they must understand to what end the tool or symbol is used. Thus, there has been a great deal of research documenting how children's understanding of adults' attentional states and intentions is crucial for language acquisition. For example, if an adult announces her intention to 'find the toma' and then searches in a number of locations, scowling upon seeing some objects and smiling upon seeing one object, children will learn the new word 'toma' for the object the adult smiles at (Tomasello and Barton 1994). In fact, Southgate et al. (2010) found that 17 month-olds learned to apply a novel word ('sefu') to a toy that an adult falsely believed was hidden in a box if the adult pointed at that box and pronounced the word (after being out of the room while the toy was moved from the box to a different location) (for a review of several similar studies, see Tomasello 1999: 114–116). Moreover, as Tomasello (1999) has emphasized, language acquisition gets going around 12 months, at which time children engage in triadic interactions with an adult and an object, and exhibit pointing behavior to inform others of events they do not know about or to share an attitude about mutually attended events others already know about (Liszkowski et al. 2007).

Apart from itself being a cultural artifact and being acquired in part by cultural learning, language enables a multitude of further effects upon cognitive development – "from exposing children to factual information to transforming the way they understand and cognitively represent the world by providing them with multiple, sometimes conflicting, perspective upon phenomena" (Tomasello 1999: 163). In acquiring a natural language, children learn to partition the world into objects and events in a way specific to their culture, and to categorize the objects and events so partitioned, and to take different perspectives upon them. Thus, one object can be described as "the dog", "fido", "the dog over there", "the golden retriever", etc., and one event can be described as "The dog bit the man", "The man was bitten by the dog", "Fido bit Daddy", etc. Which of these descriptions is appropriate depends upon the speaker's communicative goals and upon her evaluation of the listener's interests, knowledge, etc. The ability to switch among these perspectives and to deploy them flexibly is entrenched through early experiences of disagreeing with others who take different perspectives and in re-formulating utterances that have not been understood. Language also enables children to internalize rules, to memorize

information and procedures, to talk about their own reasoning processes and other experiences, and to re-describe previously implicit procedural knowledge in explicit symbolic terms, thus enabling greater flexibility and systematicity.

Finally, sensitivity to others' mental states is also of crucial importance in understanding all manner of norms that structure human sociality, since these derive their binding force not from physical facts but from agreement in people's attitudes about the statuses of entities, the entitlements and obligations they entail, etc. (Cf. Searle 1995; Gilbert 1990). And there is evidence that children as young as two are sensitive to conventional ways of doing things or using objects, and treat these conventions as normatively binding (Rakoczy et al. 2008). The relative importance of spontaneous imitation (Schmidt et al. 2010) and child-directed pedagogical cues (Gergely and Csibra 2006) is currently a matter of controversy, but both processes depend upon children's understanding an adult's goal and strategy and adopting it as the right way to perform the action in question (i.e. the way "we" do things).

9.5 Cultural Learning and the Reliability of the Intentional Stance

The next step is to close the (developmental feedback) loop by making the case that cultural learning (partially) explains the reliability of the intentional stance. One effect of cultural learning is that children become increasingly similar to the adults in their culture. More precisely, they become increasingly similar to the adults around them *as those adults appear to them on the basis of the interpretations they generate by taking the intentional stance*. This effect ensures that the use of the intentional stance during development increases its reliability as an interpretive strategy. For children's use of the intentional stance will have shaped their own development such that they themselves approximate the intentional agents that they take others to be. And, if so, they will themselves be more easily intelligible for other interpreters taking the intentional stance.

The flip-side of this, as noted briefly above (Footnote 7), is that *adults* also take the intentional stance toward young children, and that this also plays a key role in structuring children's cognitive development, i.e. by setting up expectations for them to fulfill, and by acquainting them with culture-specific objects, practices, narratives, social roles, etc. For example, gender-specific interpretations of infant behavior (such as boys' cries more often being interpreted as expressions of anger as opposed to sadness) create expectations that children then conform to (Mameli 2001). Insofar as adults' interpretation of young children as potentially rational intentional agents facilitates children's enculturation, it also increases the reliability of the intentional stance, and thus becomes, as Mameli (2001) puts it, a "self-fulfilling prophecy".

And of course this structuring effect[8] of taking the intentional stance continues into adulthood. Consider, for example, the "Knobe effect": people are more likely to interpret an action as intentional when there is a morally negative effect than when there is a morally positive effect, which can be interpreted as suggesting that one of the primary functions of intentional ascriptions is to assign blame and thereby regulate behavior (Knobe 2006). Or, more anecdotally, think of how we sanction others for departing from rational or social norms. Thus, as McGeer (2007) puts it, folk psychology is "a *regulative* practice, moulding the way individuals act, think and operate so that they become well-behaved folk-psychological agents: agents that can be well-predicted and explained using both the concepts and the rationalizing narrative structures of folk psychology" (139, emphasis in original).

Moreover, there is good reason to believe that our interpretations of our own behavior and biographies have an influence on our own actions and choices, and are thus also sometimes self-fulfilling prophecies. Some of Gazzaniga's research with split-brain patients serves as a dramatic illustration of this. One woman, whose right-hemisphere received the instruction that she should get up and leave the room, and who was then presented with a request to her left hemisphere to explain what she was doing (1995: 1393), confabulated that she had gotten up in order to get a soda – and, crucially, she then really did go and get a soda. Thus, to borrow Zawidzki's gloss on this example: "whether or not our public self-interpretations are justified or true, we actively work to confirm them" (2013: 231).

The proposal being put forward here may at first blush appear to favor simulation theory as opposed to theory,[9] since the reliability of simulations is clearly contingent upon a similarity between model and target, whereas theories do not need to be similar to whatever they explain. But in fact the proposal is neutral with respect to this dispute, because a mechanism leading to convergence of target agents and others' interpretations of them would ensure that the narratives, norms and shared understanding of objects and situations that interpreters draw upon match those that structure targets' behavior. Moreover, consider some potential sources of difficulty in everyday interpretation. Due to the holistic nature of intentional state ascription, any given interpretative act must be sensitive to a myriad of beliefs and desires that mutually constrain each other. And even if all other relevant mental states are taken

[8] Mameli (2001) coined the term "mindshaping" to denote this structuring effect of adults' intentional interpretations of children; Zawidzki (2013, Chap. 2) generalizes it to include cases, like imitative learning, where an agent actively converges upon some external model – such as models of other agents, generated by taking the intentional stance toward them.

[9] Just to recall: According to theory, social cognition is enabled by the ascription of unobservable mental states, which are defined in terms of their nomological relations to perceptions, to behavior, and to other mental states (Carruthers 2009; Gopnik 1993; Baron-Cohen 1995). Simulation theory, in contrast, is based on the idea that we generally understand others by "putting ourselves in their shoes" and using our own cognitive systems to model theirs (i.e. to simulate transitions among mental states, from perceptions to beliefs, from beliefs and desires to behavior, etc.) in which case it would (arguably) be superfluous to represent those nomological psychological relations as such (Gordon 1995; Goldman 2006; Heal 1986).

into account, they will still underdetermine the interpretation, as there is the further problem of deciding which beliefs, desires and perceptions are relevant at the moment.[10] Convergence of target agents and others' interpretations of them would help to reduce the search space and thereby increase the utility of whatever interpretive method people use, be it theory or simulation (Cf. Zawidzki 2013, Chap. 3).

The mechanism proposed here is circular, but not viciously so. It does not require that young children have the intentional states ascribed by interpreters taking the intentional stance, nor that the intentional stance is a reliable strategy when applied to very young children. It requires merely that young children *take others to have such states.*[11] If they do so, and if this provides them with role models to learn from and thereby to become more similar to, then they will develop in such a way that they subsequently become intelligible to others who also take the intentional stance. Moreover, becoming more like the model agents posited by the intentional stance also adds to children's interpretive resources, which, in turn, enable more learning and thereby more similarity, etc. In this sense, the intentional stance and cultural learning constitute a feedback loop.

9.6 The Reference of Intentional Terms

I set up this discussion of a developmental feedback loop as a response to the "If it isn't true, why does it work?" objection, and suggested that it (partially) explains the reliability of the intentional stance without appealing to intentional realism. In this section, I would like to look a bit more closely at the implications of my proposal for the realism/instrumentalism debate – and more generally, for the reference of intentional terms.

To start out by adapting one of Dennett's metaphors, what this developmental perspective encourages us to consider is that intentional states are patterns that we not only recognize and exploit but which we actively contribute to creating and sustaining. Indeed, our recognition of them is part of a causal explanation of how they are created and sustained from one generation to the next. Many patterns have this kind of dynamic structure[12]: when knitting a sweater with a pattern, one presumably monitors (and recognizes) the emerging pattern, and also modulates one's ongoing actions to ensure that the pattern is maintained. Or, if one is playing in a

[10] Cf. Dennett (1978: 125–126) on the frame problem.

[11] The effect is of course compounded by others taking the intentional stance towards them, as Mameli, McGeer and Zawidzki describe. Again, this does not require adults' intentional interpretations of children be reliable but only that they have a causal influence upon children's development.

[12] Cf. Ian Hacking's reflections on cases in which our classificatory labels (e.g. "multiple personality syndrome", "homosexuality") latch onto existing targets but also lead to changes in those targets, generating what he calls "looping effects" (e.g. Hacking 2002).

jazz band, one may respond to some perceived pattern in the music by repeating or completing it, whereupon one of the other musicians might do the same, etc.

What I would like to suggest is that this entanglement of pattern recognition, on the one hand, and pattern etiology on the other, provides an additional justification for the belief that those patterns indeed exist, because our recognition of the patterns enables us to further embed them in their respective target systems. Thus, the developmental feedback loop provides us with an additional reason to think that intentional discourse really does latch onto real behavioral patterns in the world.

Importantly, this does not require us to endorse the strong realist claim that those behavioral patterns must be mirrored by a second set of patterns that underlie behavior, i.e. that true descriptions of behavior must be isomorphic to true descriptions of the neural and/or functional processes that cause that behavior. For, although children's brains are surely molded by enculturation – and more specifically, as I am suggesting, by their use of the intentional stance to engage in cultural learning – they are molded *indirectly*. The feedback loop runs between behavioral patterns and interpretations of behavior, causing *these* to converge with each other, not the functional or neural states underlying behavior. The process must of course be enabled by functional and neural changes, but these may be just whatever changes are necessary in order to support cultural learning and thus to bring about the convergence between behavior and interpretations thereof. Thus, in order to explain why the intentional stance is a reliable strategy, it is unnecessary to postulate an isomorphism between behavioral patterns and a second set of underlying patterns that bring about those behavioral patterns.

However, as I noted above (Sect. 9.2) in discussing the "If it isn't true, why does it work?" objection, there is a lingering realist intuition that sometimes, when we see various cases as constituting a pattern, we do so by assuming some common underlying mechanism. For example: Jim *sees* an object O being placed at location L, he *hears* it being placed there, he is *told* that it is there, he acquires good reasons to *infer* that it is there, etc. We see a pattern in these cases by seeing them as instances of belief formation. Now I would like to suggest that the developmental account also makes it possible to do justice to this intuition while retaining the core Dennettian dictum that adequate intentional explanations *need not be true of the causal machinery that produces behavior.*

The basic idea is to appeal to causal theories of reference to articulate the notion that the proposed feedback loop creates a causal relationship between intentional discourse and the functional and neural mechanisms underlying behavior, and that this causal relationship provides an anchor that links the reference of intentional terms to the functional and/or neural processes that underlie behavior. Importantly, causal (as opposed to descriptive) theories of reference do not require terms or concepts to be associated with true descriptions of their referents in order to refer to them, because they maintain that reference is fixed by causal interaction with referents, not by true descriptions of them.

As a result, causal theories are better suited than descriptive theories to account for the fact that scientists (and people in general) often make referential connections despite differences in the meanings which they and others attach to terms. Similarly,

they are also better than descriptive theories in accounting for cases in the history of science where theories about some entity (e.g. the electron) have undergone dramatic change while scientists have taken themselves nevertheless to be investigating the same thing (but just to have been mistaken in their description of it).[13]

By analogy with scientists who are able to measure the effects of some novel entity and even to manipulate it in a more-or-less controlled manner prior to formulating a theory about its nature, we may think of children as tracking others' intentional states before they are capable of describing those states. They are, as it were, evolutionarily endowed with "devices" for tracking intentional states, such as eye-gaze following, emotional resonance, and mechanisms that more or less automatically detect and parse intentional action and identify goals. Although very young children have little in the way of explicit descriptions of what these "devices" track, and the descriptions that they subsequently acquire undergo various shifts over the course of cognitive development, these "devices" continue to track the same states into adulthood. Moreover, the various descriptive layers that are added on during development inherit this causal link to others' mental states and – by virtue of what I have been calling a developmental feedback loop – extend it to ever more fine-grained and more sophisticated mental states.

Thus, causal theories of reference do not require the explanations or descriptions generated by the intentional stance *to be true of* the causal machinery that produces behavior in order for them *to refer to* that machinery. This makes it possible to argue that intentional discourse refers to the real causes of behavior without correctly describing them. Thus, it does justice to the realist intuition that intentional discourse must be anchored to functional and/or neural processes, and yet at the same time it avoids the strong realist inference that adequate intentional explanations of behavior must generate true descriptions of the causal processes underlying behavior. To put this conclusion in terms of Dennett's pattern metaphor: even if the jazz musician is mistaken about the structure of the pattern that she picks up on and repeats (e.g. because she absent-mindedly thinks it consists of fourths instead of fifths), then hears it again and picks it up again etc., this iterated progression gives her confidence that she really is engaged with the pattern.

It may seems strange, after fending off the "If it isn't true, why does it work" objection, to switch back after all to a realist position. However, I submit that by retaining Dennett's commitment to the claim that adequate intentional descriptions of behavior need not be true explanations of the causal processes underlying behavior (i.e. they need not be isomorphic to adequate functional or neural explanations),

[13] It is important to note that pure causal theories are not without problems. For one thing, they make referential continuity too easy (and are thus unable to make sense of cases of referential failure), since all they require is consistency of a causal relation between agent and entity (e.g. Schouten and De Jong 1998). Moreover, there seems to be no theory-free way to pick out a natural kind or a causal entity in the first place, given that any phenomenon instantiates numerous natural kinds and many causal elements acting conjointly. A sample tiger, for instance, instantiates a species, but also a genus, and so on. The lesson seems to be that some level of description is unavoidable both in reference-fixing and reference-transmission across theory-change. Such considerations have informed more recent *causal-descriptive* theories of reference (e.g. Psillos (1999).

the causal realist option put on the table here remains compatible with the general spirit of Dennett's theory. I must emphasize, however, that causal realism is not required in order for the developmental account to provide a rebuttal to the "If it isn't true, why does it work" objection. It is simply a further theoretical option for conceptualizing the link between intentional discourse and the functional and/or neural processes underlying behavior.

9.7 Evolution, Development and Rationality

How does the developmental perspective outlined here bear upon Dennett's evolutionary response to the 'If it isn't true, why does it work?' objection? Recall that, in discussing some limitations upon Dennett's response (Sect. 9.2), I noted that we routinely ascribe beliefs and desires that depart from ideal rationality and therefore must appear aberrant from the perspective of the intentional stance theory. More specifically, we fully expect people sometimes (i) to have beliefs and desires that they ought not have, given their perceptual access, epistemic needs, biography and biological needs, (ii) to draw inferences that are not logically sound, and (iii) to act in ways that are predictable, but not rationally explicable, in light of their beliefs and desires. In discussing these cases, I also noted that Dennett has always acknowledged that the intentional stance must be supplemented and sometimes corrected with the help of empirical generalizations that people learn inductively. So I am not claiming that they undermine his position. However, adopting the developmental perspective enables us to conceptualize these supplementary resources as part and parcel of the intentional stance theory rather than as ad hoc additions: the role that the intentional stance plays in shaping cognitive development helps to account for interpreters' ability to correctly anticipate not only rational behavior but also departures from ideal rationality (as long as those departures are typical within a particular culture).

I also pointed out in Sect. 9.2 that concrete behavioral predictions are frequently underdetermined by evolutionary considerations, and that cultural knowledge is therefore required in order to reach a level of specification that is useful for predicting/explaining/influencing people's behavior in everyday life. It is rational, for example, to desire resources, but in order to bring this truism to bear in predicting an agent's behavior, it will generally be necessary to know what counts as a resource in their culture (e.g. shells, money, etc.). Note that cultural knowledge in this sense is not merely an additional tool that supplements the assumption of rationality and can be used in some range of cases where the assumption of rationality does not apply; rather, it is required in order to make use of the assumption of rationality *even in those cases where the latter does apply.* In sum, cultural knowledge acquired through cultural learning is necessary in order to specify what beliefs and desires a target agent should have *given a particular cultural context*, and to fill in gaps that an assumption of ideal rationality does not account for.

The upshot is that although evolutionary considerations constrain the determination of what intentional states to ascribe to others, it is shared developmental history that ultimately enables interpreters to fix the specific contents they ascribe to others, and to do so in ways that are fairly rational but not ideally rational. Indeed, there is likely to be an evolutionary rationale for why the intentional stance should play the shaping role that it does in development. If taking the intentional stance really does enable imitation and other forms of cultural learning, then it may support an inheritance system that has been shaped by the need to transmit behavioral phenotypes reliably from one generation to the next,[14] and by the need to increase homogeneity within human populations in order to facilitate cooperative behavior.[15] Thus, evolution may be better suited to underwrite an assumption of culture-specific *imperfect* rationality than one of *ideal* rationality.

9.8 Conclusions

I have argued that the intentional stance and cultural learning constitute a feedback loop that (partially) explains the reliability of the intentional stance, and does so – contra Dennett's realist critics – without appealing to a realist interpretation of the descriptions speakers attach to intentional terms. I have also suggested that this developmental perspective opens up the possibility of conceptualizing the link between intentional discourse and the functional and/or neural processes underlying behavior in terms of a causal theory of reference: the causal interaction between intentional interpretations of behavior and cognitive development anchors the reference of intentional terms in the functional and/or neural processes underlying behavior – and this anchoring does not require intentional explanations to be true of those functional and/or neural processes.

A further insight generated by the developmental perspective is that it is perhaps not an assumption of ideal rationality that constitutes the core of the intentional stance as an interpretive strategy but an assumption of culture-specific imperfect rationality. Interpreters expect agents to behave just as rationally as people tend to behave in their culture, and to deviate from ideal rationality in ways that are typical within their culture. Moreover, they also make specific predictions that cannot be generated simply by assuming others are (more or less) rational but which also draw upon specific cultural knowledge.

Finally, the developmental perspective outlined here also suggests a slight refinement of the criterion for determining whether to regard a target system as an intentional system. The criterion that Dennett himself has proposed is simply that the intentional stance is likely to be viable if the system in question is the product of natural selection (1978: 8). This criterion provides us with a sensible minimal requirement for applying the intentional stance, but it does not help us to understand

[14] Cf. Shea (2009).

[15] Cf. Sterelny (2003), Zawidzki (2013).

why the intentional stance works so much better when applied to other humans, especially humans with similar cultural backgrounds, than when applied to non-human animals, or to plants, bacteria and other cognitively unsophisticated evolved creatures. The developmental perspective, in contrast, enables us to account for these differences by observing that humans are particularly appropriate targets for intentional interpretation because their brains have been shaped by their own use of the intentional stance from infancy onward. This is especially true for humans with similar culture backgrounds, because the intentional stance enables children to learn culture-specific norms and practices. Thus, being the product of natural selection may constitute a minimal criterion for a system to be an apt target for intentional interpretation, but the intentional stance is likely to be especially useful when applied to a system if it is the case that cultural learning is part of the etiology of the structures that produce that system's behavior.

Acknowledgments Thanks to Mathieu Arminjon, Yanxia Feng, Ellen Fridland, Rick Griffin and Dan Dennett for stimulating discussions that helped a great deal in preparing this manuscript, and also to Gualtiero Piccinini and to two other (anonymous) reviewers for their constructive criticisms.

References

Apperly, I. (2011). *Mindreaders: The cognitive basis of theory of mind*. New York: Psychology Press.

Baillergeon, R., Scott, R., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 108–115.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*, 708–717.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Bechtel, W. (1985). Realism, instrumentalism and the intentional stance. *Cognitive Science, 9*, 473–497.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41*, 328–337.

Bischof-Köhler, D. (1991). The development of empathy in infants. In M. E. Lamb & H. Keller (Eds.), *Infant development: Perspectives from German speaking countries* (pp. 245–273). Hillsdale: Lawrence Erlbaum Associates.

Carey, S. (2009). *The origin of concepts* (1st ed.). Oxford: Oxford University Press.

Carruthers, P. (2009). How we know our minds: The relationship between metacognition and mindreading. *Behavioral and Brain Sciences, 32*, 121–182.

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B, 366*, 1149–1157.

Dennett, D. (1971). Intentional systems. *Journal of Philosophy, 68*, 87–106.

Dennett, D. (1978). *Brainstorms*. Montgomery: Bradford Press.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. (1993). Back from the drawing board. In B. Dahlbom (Ed.), *Dennett and his critics* (pp. 203–235). Oxford: Blackwell.

Dennett, D. (2008). Real patterns. In M. Bedau & P. Humphreys (Eds.), *Emergence*. Cambridge, MA: MIT Press (Originally published in: *Journal of Philosophy* 88 (1), 1991: 27–51).

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
Eisenberg, N., Spinrad, T. L., & Sadovsky, A. (2006). Empathy-related responding in children. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 517–549). Mahwah: Erlbaum.
Fodor, J. (1985). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind, 94*, 76–100.
Futó, J., Teglas, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition, 117*, 1–8.
Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge, MA: MIT Press.
Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7*(7), 278–292.
Gergely, G., & Csibra, G. (2006). Sylvia's recipe: The role of imitation and pedagogy in cultural transmission. In N. Enfield & S. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 229–255). New York: Berg.
Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415*, 755.
Gilbert, M. (1990). Walking together: A paradigmatic social phenomenon. *Midwest Studies in Philosophy, 15*, 1–14.
Goldman, A. (2006). *Simulating minds*. Oxford: Oxford University Press.
Gopnik, A. (1993). The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1–14.
Gordon, R. (1995). Simulation without introspection or inference from me to you. In T. Stone & M. Davies (Eds.), *Mental simulation: Evaluations and applications* (pp. 53–67). Oxford: Blackwell.
Griffin, R., & Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.), *Daniel Dennett: Contemporary philosophy in focus* (pp. 83–116). Cambridge, UK: Cambridge University Press.
Hacking, I. (2002). *Historical ontology*. Cambridge, MA: Harvard University Press.
Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, mind and logic* (pp. 135–150). Cambridge: Cambridge University Press.
Hoffman, M. (2000). *Empathy and moral development: Implications for caring and justice*. New York: Cambridge University Press.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130*, 203–231.
Lewis, M., & Brooks-Gunn, J. (1979). *Social cognition and the acquisition of self* (p. 296). New York: Plenum Press.
Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news and absent referents at 12 months of age. *Developmental Science, 10*(2), 1–7.
Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy, 16*, 597–628.
McGeer, V. (2007). The regulative dimension of folk psychology. In *Folk psychology re-assessed* (pp. 137–156). Dordrecht: Springer.
Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838–850.
Millikan, R. (1993). *White Queen psychology and other essays for Alice*. Cambridge: MIT Press.
Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford: Oxford University Press.
Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs N.J.: Prentice-Hall.
Prinz, J. (2005). Imitation and moral development. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation* (pp. 267–282). Cambridge: MIT Press.
Psillos, S. (1999). *Scientific realism: How science tracks truth*. London: Routledge.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Children's understanding of the normative structure of games. *Developmental Psychology, 44*(3), 875–881.
Reddy, V. (2003). On being the object of attention: Implications for self–other consciousness. *Trends in Cognitive Sciences, 7*(9), 397–402.
Richardson, R. C. (1980). Intentional realism or intentional instrumentalism. *Cognition and Brain Theory, 3*, 125–135.
Schmidt, M., Rakoczy, H., & Tomasello, M. (2010). Young children attribute normativity to novel actions without pedagogy or normative language. *Developmental Science, 14*(3), 530–539.
Schouten, M., & De Jong, H. (1998). Defusing eliminative materialism: Reference and revision. *Philosophical Explorations, 11*(4), 489–509.
Searle, J. (1995). *The construction of social reality*. New York: Free Press.
Senju, A., & Csibra, G. (2008). Gaze-following in human infants depends on communicative signals. *Developmental Science, 18*(9), 668–671.
Shea, N. (2009). Imitation as an inheritance system. *Philosophical Transactions of the Royal Society B, 364*, 2429–2443.
Song, H., Onishi, K., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected through an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition, 109*, 295–315.
Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science, 13*(6), 907–912.
Sterelny, K. (2003). *Thought in a hostile world*. Oxford: Blackwell.
Stern, D. (1985/1998). *The interpersonal world of the infant: A view from psychoanalysis and developmental psychology*. New York: Basic Books.
Stich, S. P. (1981). Dennett on intentional systems. *Philosophical Topics, 12*(1), 39–62.
Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge: Harvard University Press.
Tomasello, M., & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology, 30*, 639–650.
Tomasello, M., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences, 16*, 495–552.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 675–735.
Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science, 321*, 1831–1833.
Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy, 15*(4), 434–444.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293.
Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy through affective perspective- taking and its relation to prosocial behavior in toddlers. *Developmental Psychology, 45*(2), 534–543.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition, 13*(1), 103–128.
Yoon, J. M., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences of the United States of America, 105*(36), 13690–13695.
Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology, 28*(1), 126–136.
Zawidzki, T. (2013). *Mindshaping—A new framework for understanding human social cognition*. Cambridge, MA: MIT Press.

Chapter 10
Conscious-State Anti-realism

Pete Mandik

Abstract Realism about consciousness conjoins a claim that consciousness exists with a claim that the existence is independent in some interesting sense. Consciousness realism so conceived may thus be opposed by a variety of anti-realisms, distinguished from each other by denying the first, the second, or both of the realist's defining claims. I argue that Dennett's view of consciousness is best read as an anti-realism that affirms the existence of consciousness while denying an important independence claim.

10.1 Introduction

Philosophical discussions of phenomenal consciousness are often cast in the idiom of realism/anti-realism debates. See, for example the "phenomenal realism" discussed by Chalmers (2003), Block (2002), and McLaughlin (2003) as well as the "qualia realism" discussed by Kind (2001), Graham and Horgan (2008), and Hatfield (2007). Often, the realists label themselves as such in the interest of making an existence claim and casting their opponents as those nihilists or eliminativists who would deny the existence of phenomenal consciousness and/or qualia. For example, critics of Daniel Dennett often characterize him as denying the very existence of consciousness.[1] But, at least sometimes, there is more being claimed by the realists than the mere existence of consciousness: They are claiming that what exists also exists *independently* (Independently of what? More on this shortly). It's open, then, for a consciousness anti-realist to affirm the existence of consciousness while denying that its existence is independent in a way interesting to realism/anti-realism debaters. My aim in the present paper is to explore such an existence-affirming consciousness anti-realism, especially as exemplified in Daniel Dennett's career-spanning work on consciousness, key components of which of course include his

[1] See, for example, Strawson (2009, pp. 51–52), Searle (1997, p. 120), Block (1997, p. 75), Seager (1999, p. 85).

P. Mandik (✉)
William Paterson University of New Jersey, Wayne, NJ, USA
e-mail: petemandik@gmail.com; mandikp@wpunj.edu

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_10

books *Content and Consciousness* (1969) as well as *Consciousness Explained* (1991b), and *Sweet Dreams* (2005).

What sense can be made of independence in the context of discussions about consciousness? In other realism debates—debates, for instance, about numbers, colors, or physical objects—independence claims are often cast in terms of mind-independence (Khlentzos 2011). A realist about electrons holds that electrons would still have existed even if no minds did. A realist about colors holds that an object can have a color even if no mind exists to perceive its color. While formulations of independence claims along such lines may make sense for colors and physical objects, they may initially seem ill-suited for making coherent independence claims about phenomenal consciousness. It makes little sense to say that consciousness could have existed even if no minds existed. It makes little sense to say that qualia exist independently of how things are perceived or experienced. Despite the inapplicability of these forms of independence claims to consciousness, there is a sensible way of interpreting a relevant independence claim: It is a claim about consciousness occurring independently of what one *thinks* or *believes*. The anti-realism under present consideration denies this sort of independence claim.

The consciousness anti-realism I focus on in the present paper is a view that Dennett has defended across several works—it's part of the "semi realism" of his "Real Patterns" (1991a), a view of consciousness described by Dennett (1994) as opposed to "hysterical realism". Given the way that Block (2002, p. 392) characterizes "phenomenal realism" as a thesis that "allows the possibility that there may be facts about the distribution of consciousness which are not accessible to us even though the relevant functional, cognitive, and representational facts are accessible," Dennett may appear to certain eyes to be an anti-realist merely for the fact that his view on consciousness dating all the way back to *Content and Consciousness* is functionalist, cognitivist, and representationalist. However, there is a much more specific anti-realist view of Dennett's that I want to focus on here. In *Consciousness Explained*, Dennett describes this view as "first-person operationalism," a thesis that "brusquely denies the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness" (1991b, p. 132).

Dennett's most famous argument for his first-person operationalism (hereafter, FPO) proceeds by pointing out the alleged empirical underdetermination of theory-choice between "Stalinesque" and "Orwellian" explanations of certain temporal anomalies of conscious experience (Dennett, op. cit., pp. 115–126). The explanations conflict over whether the anomalies are due to misrepresentations in memories of experiences (Orwellian) or misrepresentations in the experiences themselves (Stalinesque).

David Rosenthal (1995, 2005b, c) has offered that his Higher-order Thought theory of consciousness (hereafter, "HOT theory") can serve as a basis for distinguishing between Orwellian and Stalinesque hypotheses and thus as a basis for resisting FPO. The gist of HOT theory is that one's having a conscious mental state consists in one's having a higher-order thought (a HOT) about that mental state.[2]

[2] Such a HOT must also not be apparently arrived at via a conscious inference, but this further constriction on the HOTs that matter for consciousness is of little importance to the present paper.

I'll argue that HOT theory can defend against FPO only on a "relational reading" of HOT theory whereby consciousness consists in a relation between a HOT and an actually existing mental state. I'll argue further that this relational reading leaves HOT theory vulnerable to objections such as the Unicorn Argument (Mandik 2009). To defend against such objections, HOT theory must instead admit of a "nonrelational reading" whereby a HOT alone suffices for a conscious state. Indeed, HOT theorists have been increasingly explicit in emphasizing this nonrelational reading of HOT theory (Rosenthal 2011; Weisberg 2010, 2011). However, I'll argue, on this reading HOT theory collapses into a version of FPO.

The remainder of the paper will go like this: In Sect. 10.2 I'll say some more about Dennettian anti-realism (FPO) and the Orwellian/Stalinesque argument. In Sect. 10.3 I'll lay out a HOT-theoretic version of the Orwellian/Stalinesque distinction that depends on a relational reading of HOT theory. In Sect. 10.4 I'll spell out the case for a nonrelational reading of HOT theory and how HOT theory is thereby led to a kind of FPO.

10.2 Anti-realism, Consciousness, and FPO

10.2.1 Clarifying Consciousness Anti-realism

In this subsection I want to rapidly clarify key terms. My aim in the present section is not to argue that one set of construals is better than another, but instead to lay out a series of stipulations to facilitate the rest of the discussion.

Consciousness aside for a moment, let's think about the general structure of realism/anti-realism theses and debates between them. A realist position, say realism about dogs, is a conjunction of an existence claim and an independence claim, where the independence in question is often glossed as "mind independence". An imprecise statement of dog realism is "dogs exist and exist mind-independently." Each conjunct admits of multiple precisifications. I'll have little to say in the present paper about precisifications of the existence claim. Let it suffice that I intend existence claims to be tenseless and actual-world directed. So, items in the past and future exist, though no item in a nonactual possible (or impossible) world does. The extinction of dogs will not, then, falsify dog realism.

Precisifications of the independence claims require more care, especially if we want to formulate coherent claims of mind-independence about things that are themselves mental. One precisification of independence that will not serve present purposes is one stated simply in terms of minds, as in "X exists independently of any mind existing." Clearly, plugging "minds" in for "X" generates an incoherence. Precisifications that avoid such an incoherence appeal instead to specific kinds of mental state, say specific kinds of thought, belief, or judgment. "Minds exist independently of anyone thinking, believing, or judging that minds exist" contains no obvious incoherence. Precisifications of the independence claim along this line will be what I have in mind for the rest of the paper. Of interest will be the question of

whether one's conscious experience exists independently of one's thinking, believing, or judging it to exist.

Given that realist theses are each a conjunction of an existence claim and an independence claim, opponents of realism come in two varieties: Nihilists, who deny the existence claim, and idealists, who deny the independence claim. A Berkeleyan idealist about dogs (a "bark"-leyan?) does not deny that dogs exist, but instead denies that dogs exist independently of being perceived.

I will simply set nihilism aside in this paper, and reserve "anti-realism" for the idealist variety. While Dennett's critics sometimes accuse him of denying that consciousness exists, it should be clear that Dennett's statement of FPO doesn't support such a reading. In denying "the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness," Dennett is clearly not denying an existence claim, but instead an independence claim. The kind of anti-Dennettian that I am interested in can be briefly described as holding that we can sort mental states into two varieties, experiences and thoughts, and that conscious instances (and facts about them) of the first variety obtain independently of instances of the second variety.

One further set of issues I want to address before leaving this subsection concerns which facts about consciousness are at issue. What we get directly from the Dennett quote is that FPO is anti-realist about "consciousness of a stimulus". Some consciousness theorists, especially HOT theorists, will detect an ambiguity in this phrase. Many, if not all, follow Rosenthal in distinguishing "transitive consciousness" (being conscious of something) from "state consciousness" (a mental state's being conscious) (For Rosenthal's discussion of the distinction, see, for instance, (2005a, p. 4)). If there is such a distinction, then the possibility opens of having a state in virtue of which one is conscious of something without that state itself being a conscious state. For example one might have a perceptual state by which one is conscious of a red rose without the perceptual state itself being conscious. Other theorists do not urge such a distinction. Dretske, for instance, says that conscious states are states "we are conscious *with*, not states we are conscious *of*" (1995, pp. 100–101). Perhaps (though I'm unsure) Dennett counts among such theorists. However, regardless of where one stands on this issue, there is an interesting anti-realist thesis to be stated explicitly in terms of state consciousness. Modifying the Dennett quote accordingly yields a thesis that "brusquely denies the possibility in principle of *a conscious experience of a stimulus* in the absence of the subject's belief in that consciousness" (altered text italicized). For the remainder of the paper, I shall be interpreting FPO as including this thesis.

Before proceeding to the next section, I should note that, contra Kiefer (2012a) and Muñoz-Suárez (personal correspondence) one view of Dennett's that is *not* a part of FPO is his view that certain of a speaker's speech acts determine the contents of that speaker's intentional states. This thesis of a dependence of thought upon speech and other expressions is separable from FPO, which is a thesis of a dependence of consciousness upon thought.

10.2.2 The Orwellian/Stalinesque Argument for FPO

The phi phenomenon is a species of illusory motion, as when one views the flashing stationary lights on a marquee. Color phi is a species of the phi phenomenon in which the stationary stimuli differ in color and the apparently moving object changes color mid-trajectory. Subjects in a color phi experiment look at a computer screen upon which a green circle appears then disappears. A small time later in a position a small distance away from where the green circle was, a red circle of the same size appears and then disappears. The time elapsed between the disappearance of the green and the appearance of the red is very short. It's so short that, as a subject in this experiment, it would appear to you as if a single circle appears, moves across the screen, and then disappears. Further, the single moving circle would appear to start off green and change to red midway in its trajectory. This is color phi and it is weird.

Color phi is not just weird because we don't know how the brain creates illusory motion from nonmoving stimuli. Here's the really puzzling thing about color phi: How does the brain know to change the moving green circle to red *before* the red circle appears? Clairvoyance aside, clearly it cannot. So the experience of the red-to-green change needs to have happened after the brain receives information of the appearance of the red circle. We want further details in explaining this, and here we feel pulled toward two competing explanations, explanations that Dennett famously dubs "Orwellian" and "Stalinesque".

My mnemonic for Dennett's labels is that "Stalinesque" shares an "s" and a "t" with "show trial," and "Orwellian" has an "r" in common with "revisionist history." Both explanations have key roles for the notions of consciousness and of falsehood, but differ with respect to the questions of which states are conscious and which ones are false representations.

Let's start by looking at the revisionist history, that is, the false memory, posited by the Orwellian explanation. On this explanation, the key mental events and their temporal order are as follows: First there is a conscious experience of a green circle, next there is a conscious experience of a red circle, and finally there is a false memory of a single circle having moved and changed from green to red. On the Orwellian explanation, there is neither a conscious experience of motion nor one of color change, but instead a false memory that movement and color change were experienced.

Let us turn now to the Stalinesque explanation, which posits a show trial. On this explanation, the false mental state posited is not a memory but an experience. On the Stalinesque explanation, the key mental events and their temporal order are as follows: First there is an unconscious receipt of information concerning the green circle, next there is an unconscious receipt of information concerning the red circle, and finally, based on these raw materials, a conscious experience is assembled—a false experience of a green circle moving and changing to red mid-trajectory.

On the face of it, these seem to be distinct competing explanations of the empirical data. The Orwellian explanation posits two accurate conscious experiences of two stationary, differently colored circles followed by a false memory of having experienced a single moving circle that changes color. The Stalinesque explanation posits a false conscious experience of motion and mid-trajectory color-change and an accurate memory of that experience. To highlight their differences, we can describe the explanations as follows: the Orwellian posits a false memory and accurate conscious experience, whereas the Stalinesque posits a false conscious experience and an accurate memory (of what the experience was).

If these are indeed distinct explanations, then which one is the correct one? Dennett argues persuasively that no amount of evidence, either first-personal or third-personal, will determine theory choice here. I'm persuaded. I find it easy to be so persuaded.

To attempt to persuade yourself of Dennett's conclusion, first imagine being a subject in a color phi experiment. What you introspect is that there has been a visual presentation of a moving, color-changing circle. Your introspective judgment is that you have experienced such an episode. But to resolve the Stalinesque v. Orwellian debate on introspective grounds, your introspective judgment would need to wear on its sleeve whether its immediate causal antecedent was a false memory (Orwellian) or a false experience (Stalinesque). But clearly, no such marker is borne by the introspective judgment. So much for the first-person evidence!

So now, imagine being a scientist studying a subject in a color phi experiment. Imagine availing yourself of all of the possible third-personal evidence. Suppose you avail yourself to evidence gleaned via futuristic high-resolution (both spatially and temporally) brain scanners. Such evidence, let us suppose, will allow you to determine not only which brain events occur and when, but also which brain events carry which information, and which brain events are false representations. This is, of course, to presume solutions to very vexing issues about information, representation, and falsehood, solutions that might beg the question against a Dennettian anti-realism about representation and perhaps, thereby, against Dennettian anti-realism about consciousness, but I won't pursue this line of thought here. However, we will here suppose that such solutions can be arrived at independently of resolving issues about consciousness. Clearly, then, the evidence that you have will, by itself, tell you nothing about which states are conscious. So much for the third-person evidence!

To surmount this hurtle for strictly third-person approaches, you may feel tempted to either ask the subject what their conscious experiences are like, or allow yourself to be a subject in this experiment. However, either way you will only gain access to an introspective judgment with a content that we have already seen as underdetermining the choice between the Orwellian and the Stalinesque.

Given that there's no real difference between the Orwellian and Stalinesque scenarios, what matters for consciousness is what the scenarios have in common, namely the content of the belief or thought that one underwent a conscious experience of a color-changing, moving circle. There's nothing independent of this belief content that serves to make it true, so having a belief with such-and-such content is all there is to being in so-and-so conscious state.

10.3 HOT Orwellian and HOT Stalinesque Scenarios

Dennett's Orwellian/Stalinesque argument turns on a kind of underdetermination of theory by evidence. Of course, what evidence underdetermines, additional theory can sometimes settle. Rosenthal constructs HOT-theoretic versions of the Orwellian and Stalinesque scenarios that are distinguishable given the resources of HOT theory (1995, p. 362). However, that there are *some* Orwellian and Stalinesque scenarios that are distinguishable from each other doesn't suffice to refute FPO. Dennett himself admits that some Stalinesque scenarios are distinguishable from some Orwellian scenarios (especially at macroscopic time-frames) (Dennett 1991b, p. 117). What matters instead is that there are some Orwellian and Stalinesque scenarios that are not distinguishable from each other. I aim in the present section to show that there are Orwellian and Stalinesque scenarios that HOT theory serves to distinguish only on a relational reading of HOT theory.

One way to convey the gist of HOT theory is by saying that a state is conscious when a HOT is about that state. Reading this relationally, we have two relata and a relation between them. The relata are the HOT and the state that it is about. The relation the HOT bears to its target is an "aboutness" relation, or as I'll prefer to say, a "representing relation". So, when a visual experience of a red circle is accompanied by a HOT that bears the representing relation to it, then the visual experience is a conscious one. If, instead, the visual experience is unaccompanied by any such HOT, the experience is an unconscious one. Sometimes HOT theorists themselves put HOT theory in ways that invite the relational reading. For example, Rosenthal (2005c, p. 322) writes that his is "a theory according to which a mental state is conscious just in case it is accompanied by a higher-order thought (HOT) to the effect that one is in that state." Prima facie, this talk in terms of accompaniment makes a representing relation seem central to HOT theory. However, perhaps in the final analysis Rosenthal's commitment to the relational reading may be merely a superficial appearance. I'll return to this issue in Sect. 10.4. For the present section, I will keep the relational reading at the forefront.

With this relational reading of HOT theory in mind, let us think through how color phi can be explained. In color phi, it seems to one that one has an experience of a moving circle that changes color. In order for it to seem to one that one is having an experience of a moving, color-changing circle, there needs to be a HOT that one is visually experiencing a moving, color-changing circle. We might wonder further about what the causal antecedents are of this HOT, especially as concerns links in the causal chain after the information from the stationary flashing circles has hit the eye of the beholder.

One possibility is that none of the causal antecedents of the HOT is a visual experience of motion and color change. Instead, the causal antecedents are visual experiences of the stationary red and green circles. Further, it is a consistent elaboration on this possible scenario that no causal consequence of the HOT is a visual experience as of motion and color change. Since nothing antecedent or consequent to the HOT answers to the description that constitutes the HOTs content, the HOT

is false. Since the HOT is not itself an experience (it is instead a thought) and has occurred after the experiences that triggered its occasion, we can regard it as a memory (albeit, a false one). Given the possibility we've just consistently described, this reading of the HOT theory casts it as close to Orwellian. However, to be fully Orwellian, there needs to be posited, in addition to a false memory, an accurate conscious experience. Can we complete an Orwellian explanation sketch that is consistent with HOT theory? I think that we can, but some care needs to be taken.

The way to introduce an accurate conscious experience into the above sketch in a way that is consistent with HOT theory is to go looking for one or more states that the HOT is about. If this sketch is to be Orwellian, some choices for what the HOT is about will be better than others. On a highly natural reading of what the HOT is about, it is about an inexistent state, namely a visual experience of motion and color change. The inexistence of such a state is what makes the HOT false. One problem with this reading is that the Orwellian is supposed to be positing the *existence* of a conscious state, and it is highly strained to posit the existence of something that is admitted in the same breath to not exist. I hope I will be forgiven in dismissing the Meinongian perspective required to view existing inexistents as welcome company. Anyway, there is another problem: It is difficult to regard the inexistent state as accurate. The inexistent state is a representation of movement and color change upon the computer screen, and, in actuality, no such motion or color change exists. And since Meinongianism is here not taken seriously, there is no serious way of taking the suggestion that the inexistent state is an accurate representation, albeit one that accurately represents an inexistent state of affairs.

There is another possibility for interpreting what the HOT is about, namely that it is about the two separate experiences of the differently colored circles. In being about those accurate experiences, they are thereby rendered conscious: On the occasion of the HOT about them, the experiences become conscious. This may have a slight air of strangeness, but there's no obvious problem in a representation of something representing it falsely. Indeed, the scenario described here is a possibility that Rosenthal explicitly endorses (2005b, pp. 240–241) (That is, he endorses it as a possibility. He does not assert that it is an actuality).

Thus completes my sketch of a HOT Orwellian explanation of color phi. Let's try to fit a Stalinesque explanation into the HOT mold as follows: Recall that a Stalinesque explanation posits a false conscious experience of motion and mid-trajectory color-change that has as causal antecedents the unconscious receipt of information concerning the stationary presentations of the green circle and the red circle. To fit such an explanation into the HOT mold, the HOT theorist needs to posit a HOT that is about an experience that is itself (the experience) a false representation of motion and color change. Otherwise, without such a HOT, the false experience won't be conscious. But in order to introduce this HOT, a means must be devised of determining that the HOT is about the false representation and not about the accurate representations. Otherwise, the accurate representations will be the conscious ones and the proposed explanation won't be Stalinesque. Supposing that such a means can be determined, we therefore have a Stalinesque reading of a HOT-theoretical explanation of color phi.

It looks, at least prima facie, that HOT theory is consistent with Orwellian and Stalinesque explanations. However, once these explanations are fit into the HOT mold, are opportunities thereby made available for adjudicating between them?

Note the key similarities in the Orwellian and Stalinesque stories. On both stories there is a HOT, the content of which is that there's an experience of motion and color change. Also, on both stories there are accurate experiences of the stationary red and green circles. The key differences are that, on the Orwellian story, the HOT bears the representation relation to the accurate experiences and not to the (inexistent) inaccurate experience of motion and color change. On the Stalinesque story, the HOT bears the representation relation to the inaccurate experience of motion and color change and not to the accurate experiences of the stationary red and green circles. If we assume that the HOT theory is true, then in order to discover whether color phi is Orwellian or Stalinesque we would need to discover whether the HOT bore a representing relation to the accurate experiences or not.

To give a preview of the worry that I ultimately want to press against HOT theory, there are good reasons to think that there is no such thing as a representation relation and so, if the HOT theory is true, no such relation figures in it. But without recourse to such a relation, there is no relevant difference between the HOT Orwellian and the HOT Stalinesque explanations: On either case, the content of one's consciousness just is the content of the HOT, and that content is the same on either story.

10.4 Non-relational HOT Theory and FPO

Elsewhere I press an argument, "the Unicorn Argument" or just "the Unicorn," against HOT theories (Mandik 2009). At the heart of the argument is a view about how best to think of representation in the face of the representation of inexistents such as unicorns. This view can be seen as emerging as a response to the famous inconsistent triad of intentionality.[3] One way of presenting the triad is like this:

1. Representing is a relation borne to that which is represented.
2. There are representations of inexistents.
3. There are no relations borne to inexistents.

While all three propositions of the triad are independently plausible, they cannot be jointly true. The heart of the Unicorn involves a denial of the first item in the triad while retaining the last two. The resulting view might be summed up as holding that there is no such thing as a representing relation—representation may involve relations, but it is not constituted by a relation to that which is represented. It follows from there being no representation relation that there is no such relational property as the property of being represented.

[3] For further discussion of the inconsistent triad of intentionality see Crane (2001, especially pp. 22–28), Kriegel (2007, especially pp. 307–312, 2008), and Mandik (2010, p. 64, 2013, p. 188).

This line of thought is pressed against the HOT theory by reading HOT theory as committed to the existence of such relations and relational properties. On what I'll call the "relational reading" of HOT theory, a state is conscious only if a HOT bears the representing relation to that state. On this reading of HOT theory, the property of being conscious just is the property of being represented by a HOT. Read relationally, HOT theory gives a nicely straightforward explanation of how one and the same mental state can be unconscious at one time and conscious at another time. The change from being unconscious to being conscious just is the change from not being appropriately related to a HOT to being so related. And what is this relation if not a representing relation?

For examples of theorists who interpret HOT theory along such relational lines see Gennaro (2006, 2012), Wilberg (2010), and Bruno (2005). For discussions of both relational and non-relational interpretations of HOT theory, see Lau and Brown (n.d.), Brown (2012), Berger (2013), and Pereplyotchik (2015). What's relational about the relational reading is the required existence of an actual state for the HOT to be about. One is in a conscious state, when a HOT bears a certain sort of relation toward another mental state, M. The relation borne to M is presumably that the HOT represents or is about M. On, for example, Gennaro's view, M and the HOT are held to be proper parts of a mereological fusion and the fusion is the conscious state. Nonetheless, even on Gennero's view, a key role is played by the HOT's relating to M by way of an aboutness or a representing relation.

However, and this is the thrust of the Unicorn, if there are no such relations (as the representing relation) and relational properties (as the property of being represented), and there *is* such a property as a state's being conscious, then being represented cannot be what a state's being conscious consists in.

Some HOT theorists often present their view in a way that seems to invite the relational reading. However, in responding to the Unicorn and closely related objections turning on "empty" higher thoughts (e.g. Byrne 1997; Neander 1998; Block 2011), some HOT theorists have urged a reading of their view that I'll call the "non-relational reading."[4]

Weisberg (2010), in responding to the Unicorn, cites approvingly a remark of Harman's (1997), part of which includes the statement "I am quite willing to believe that there are not really any nonexistent objects and that apparent talk of such objects should be analyzed away somehow" (p. 423, fn. 26). Rosenthal (2011) writes, in response to Block's (2011) attack based on empty HOTs:

> Block describes me as having retreated from an 'aboriginal' theory, on which the targets of HOTs always exist, to a 'new version' on which they need not […]. This is not so; in my earliest publication about consciousness I noted the possibility of absent first-order states […]. For ease of exposition, I often introduce the theory by saying that a state is conscious when it's accompanied by a HOT, noting that this characterization is not strictly accurate.

[4] Alex Kiefer (2012a, b) suggests that even on this non-relational reading of HOT theory, sufficient sense can be made of an existent first-order state's being represented by an accurate higher-order thought. I worry, however, that the suggested proposal cannot be spelled out without problematically quantifying into the opaque context introduced by the relevant higher-order thought.

> And there's no harm in putting things in those relational terms when the existence of HOTs' targets is not under consideration.
>
> All that matters for a state's being conscious is its seeming subjectively to one that one is in that state. On the HOT theory, that's determined by a HOT's intentional content […]. (p. 436)

With this nonrelational reading of HOT theory in mind, it becomes overwhelmingly difficult to see how HOT theory isn't just a version of FPO. In publications attacking FPO, Rosenthal describes FPO as, among other things, a view whereby "facts about…when states become conscious are exhausted by how things appear to consciousness" (Rosenthal 2005c, p. 323). Note how similar such a description of FPO is to Rosenthal's own description of HOT theory in publications highlighting its invulnerability to empty-HOT based attacks: "A state's being conscious is a matter of mental appearance—of how one's mental life appears to one" (Rosenthal 2011, p. 431). The core similarities between FPO and nonrelational HOT theory are (1) a state's being conscious is its appearing to one that one is in such-and-such mental state, and (2) the relevant way in which one is appeared to is via thought—it appears to one that one is in such-and-such mental state when one *thinks* (as opposed to senses or imagines) that one is in such and such mental state.

I find it hard to shake the impression that there is a tension within HOT theory itself between a relational reading and a nonrelational reading. Further it seems that the nonrelational reading is highlighted when defending against empty-HOT and Unicorn types of objections and that the relational reading is highlighted when defending against FPO. In a publication targeting FPO Rosenthal (1995) seems himself to be promoting a relational reading of HOT theory:

> Because many mental states aren't conscious at all, it's implausible that the property of being conscious is an intrinsic property. All mental states have some sort of content properties—intentional content in the case of intentional states and sensory content in the case of bodily and perceptual sensations and most emotions. Such content properties are arguably intrinsic to mental states. By contrast, mental states can be conscious at one moment and not at another; so we have no reason to regard the property of being conscious as being intrinsic to such states. Accordingly, a state's being conscious requires the occurrence of something extrinsic to it. And it may well be, therefore, that no mental state is conscious when it first occurs. But this doesn't mean there are no facts of the matter about consciousness; states are conscious when, and only when, the relevant events occur. (p. 364)

Describing the requirements on a state's being conscious in terms of "the occurrence of something extrinsic to it" points quite strongly in the direction of the relational reading of the HOT theory. There is posited here a key role for a relation between two states: the conscious state and the HOT that is about that state. And this is in clear tension with the non-relational reading that seems most naturally applicable to the insistence, in Rosenthal (2011), that the HOT all on its own suffices for state consciousness and there being something it's like.

If there is a way to resolve the apparent tension between relational and nonrelational readings of HOT theory, I do not know what it is. I do hope, though, that the present paper aids in progress toward a resolution. It has been my aim in this paper to argue that the HOT theory can be defended as an alternative to Dennett's FPO only by reading HOT theory as a relational theory. It seems to me, however that the

balance is tipped toward a nonrelational reading of HOT theory and thus, if my arguments are correct, a reading of HOT committing it to FPO.

Acknowledgements I am especially grateful to Alex Kiefer and Josh Weisberg for detailed discussions of earlier versions of the present work. I am also grateful for feedback from and discussions with James Dow, Jacob Berger, Richard Brown, Aspasia Kanellou, Daniel Kostic, Carlos Muñoz-Suárez, Adriana Renero, David Rosenthal, Miguel Sebastian, and Josh Shepherd. I am grateful for the audiences of presentations of this material at the University of Houston Department of Philosophy colloquium and the Fourth Annual Online Consciousness Conference.

References

Berger, J. (2013). Consciousness is not a property of states: A reply to Wilberg. *Philosophical Psychology*, 1–21. doi:10.1080/09515089.2013.771241

Block, N. (1997). Begging the question against phenomenal consciousness. In N. Block, O. Flanagan, & G. Guzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 175–180). Cambridge, MA: MIT Press.

Block, N. (2002). The harder problem of consciousness. *Journal of Philosophy, 99*(8), 391–425.

Block, N. (2011). The higher order approach to consciousness is defunct. *Analysis, 71*(3), 419–431. doi:10.1093/analys/anr037.

Brown, R. (2012). Review of Rocco J. Gennaro the consciousness paradox: Consciousness, concepts, and higher-order thoughts. *Notre Dame Philosophical Reviews*, 1–7. Retrieved from http://ndpr.nd.edu/news/30848-the-consciousness-paradox-consciousness-concepts-and-higher-order-thoughts/

Bruno, M. (2005). A review of Rocco J. Gennaro (ed.) Higher-order theories of consciousness: An anthology. *Psyche, 11*(6), 1–11.

Byrne, A. (1997). Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies, 86*, 103–129.

Chalmers, D. (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical essays* (pp. 220–272). Oxford: Oxford University Press.

Crane, T. (2001). *Elements of mind: An introduction to the philosophy of mind*. Oxford: Oxford University Press.

Dennett, D. (1969). *Content and consciousness*. London/Boston/Henley: Routledge & Kegan Paul.

Dennett, D. (1991a). Real patterns. *Journal of Philosophy, 88*, 27–51.

Dennett, D. (1991b). *Consciousness explained*. Boston: Little Brown and Company.

Dennett, D. C. (1994). Get real. *Philosophical Topics, 22*(1&2), 505–568.

Dennett, D. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Gennaro, R. (2006). Between pure self-referentialism and the (extrinsic) HOT theory of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 221–248). Cambridge, MA: MIT Press.

Gennaro, R. (2012). *The consciousness paradox: Consciousness, concepts, and higher-order thoughts*. Cambridge, MA: MIT Press.

Graham, G., & Horgan, T. (2008). Qualia realism, its phenomenal contents and discontents. In E. Wright (Ed.), *The case for qualia* (pp. 89–107). Cambridge, MA: MIT Press.

Harman, G. (1997). The intrinsic quality of experience. In N. Block, O. J. Flanagan, & G. Guzeldere (Eds.), *The nature of consciousness* (pp. 663–675). Cambridge, MA: MIT Press.

Hatfield, G. (2007). The reality of qualia. *Erkenntnis, 66*(1), 133–168.

Khlentzos, D. (2011). Challenges to metaphysical realism. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/spr2011/entries/realism-sem-challenge/

Kiefer, A. (2012a, February 17). Comments on Pete Mandik's 'Conscious-state Anti-realism'. *Consciousness Online, 4*. Retrieved from https://consciousnessonline.files.wordpress.com/2012/02/kiefer-comments-on-mandik.pdf

Kiefer, A. (2012b, March 23). *Higher-order representation without representation*. 104th Annual Meeting of the Southern Society for Philosophy and Psychology, Savannah.

Kind, A. (2001). Qualia realism. *Philosophical Studies, 104*(2), 143–162.

Kriegel, U. (2007). Intentional inexistence and phenomenal intentionality. *Philosophical Perspectives, 21*, 307–340.

Kriegel, U. (2008). The dispensability of (merely) intentional objects. *Philosophical Studies, 141*, 79–95.

Lau, H., & Brown, R. (n.d.). The emperor's new phenomenology? The empirical case for conscious experience without first-order representations. In A. Pautz & D. Stoljar (Eds.), *Festschrift for Ned Block*. Cambridge, MA: MIT Press.

Mandik, P. (2009). Beware of the unicorn: Consciousness as being represented and other things that don't exist. *Journal of Consciousness Studies, 16*(1), 5–36.

Mandik, P. (2010). *Key terms in philosophy of mind*. New York: Continuum.

Mandik, P. (2013). *This is philosophy of mind: An introduction*. Oxford: Wiley-Blackwell.

McLaughlin, B. P. (2003). A naturalist-phenomenal realist response to Block's harder problem. *Philosophical Issues, 13*(1), 163–204.

Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. *Nous, 32*(S12), 411–434.

Pereplyotchik, D. (2015). Some HOT family disputes: A critical review of *The Consciousness Paradox* by Rocco Gennaro. *Philosophical Psychology, 28*(3), 434–448.

Rosenthal, D. (1995). Multiple drafts and facts of the matter. In T. Metzinger (Ed.), *Conscious experience* (pp. 359–372). Exeter: Imprint Academic.

Rosenthal, D. (2005a). *Consciousness and mind*. Oxford: Clarendon Press.

Rosenthal, D. (2005b). First-person operationalism and mental taxonomy. In D. Rosenthal (Ed.), *Consciousness and mind* (pp. 229–256). Oxford: Oxford University Press.

Rosenthal, D. (2005c). Content, interpretation, and consciousness. In D. Rosenthal (Ed.), *Consciousness and mind* (pp. 321–335). Oxford: Oxford University Press.

Rosenthal, D. (2011). Exaggerated reports: Reply to Block. *Analysis, 71*(3), 431–437. doi:10.1093/analys/anr039.

Seager, W. (1999). *Theories of consciousness: An introduction and assessment*. New York: Routledge.

Searle, J. (1997). *The mystery of consciousness*. New York: Review Books.

Strawson, G. (2009). *Mental reality* (2nd ed.). Cambridge, MA: MIT Press.

Weisberg, J. (2010). Misrepresenting consciousness. *Philosophical Studies, 154*(3), 409–433. doi:10.1007/s11098-010-9567-3.

Weisberg, J. (2011). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis, 71*(3), 438–443. doi:10.1093/analys/anr040.

Wilberg, J. (2010). Consciousness and false HOTs. *Philosophical Psychology, 23*(5), 617–638.

Chapter 11
Not Just a Fine Trip Down Memory Lane: Comments on the Essays on *Content and Consciousness*

Daniel Dennett

Abstract The current chapter contains commentaries and replies to all nine essays included in the present volume.

It has been more than a pleasure to read and reflect on these thoughtful, constructive essays, which have taught me a lot about my own work—always a bracing experience—and pointed to future directions well worth exploring further. I think what makes me most proud of my firstborn book (published when I was 27-years-old) is that 45 years later it can still provoke high caliber work like this. These essays are not backward-facing nostalgic reflections on an antique book of merely historical interest but forward-looking appropriation and exploitation of ideas that are useful, their authors think, on the cutting edge today. It is never gracious to say "I told you so," but sometimes the urge to say it is strong, so it is particularly gratifying to have these excellent philosophers say it for me, in nine different ways. I don't agree with everything they have to say, but where we still disagree, I may well be missing something they understand better than I do.

1

If **Don Ross** is right, it is possible to make a major scientific discovery without trying, and without recognizing that you've done so. I don't view that conditional as an invitation to perform *modus tollens*. It is quite possible to stumble into something more important than you realize at the time, and Ross suggests as much:

> It is instructive that of all the articulations of his theory of the mind that Dennett has produced over the course of his career, the one most strongly based in traditional philosophical analysis and argumentation got the science of the story right in all its essentials – as judged against both Dennett's later opinions and what has been implicitly endorsed by later scientific practice – while providing an unstable and unsatisfactory account of the metaphysics. (p. 37)

D. Dennett (✉)
Tufts University, Medford, MA, USA
e-mail: daniel.dennett@tufts.edu

C. Muñoz-Suárez, F. De Brigard (eds.), *Content and Consciousness Revisited*,
Studies in Brain and Mind 7, DOI 10.1007/978-3-319-17374-0_11

I "got the science right" almost by accident, while engaging in "traditional philosophical analysis and argumentation." And the metaphysics has been "unstable." Fair enough. My initial forays into the scientific literature were the efforts of an utter novice with no scientific training, but it does seem that I had a knack for finding the best scientific beacons—and interpreting them in a philosophically novel way. And, truth to tell, at the time I was working in very solitary fashion on my dissertation and later on its descendant, *C&C,* it did *seem to* me that I was opening up some productive new vistas, that I had found a way out of some perennial *philosophical* traps by taking seriously some ideas I had developed about how to understand the scientific project. But that hunch also often seemed too good to be true. I worried that I must be missing something that others understood, something that explained why what seemed to me like exciting new ideas were, in the end, forlorn. And to be sure, there are quite a few eminent philosophers who are utterly confident that I *am* missing something, and have been saying so for years, but at this point, I'm much more sure of my ground. From my perspective, *C&C* was a lucky strike of a mother lode of ideas that, with a little refining, form pieces that snap together in a very productive way. So I think Ross has nicely uncovered and "celebrated," as he says, something of a discovery that I more-or-less made back in the 60s and have been trying to understand better and defend ever since. As he notes, I have not always had the best version of my position in focus, and over the years he has intervened on occasion to help me get back on the right track, as he sees it. I don't always agree with his proposed improvements, some of which I may not understand, but since he's often been dead right, I take all his suggestions seriously.

Ross speaks of the "over-reaction against behaviorism," which was certainly a hallmark of the early days of cognitive science, largely inspired by Chomsky's caricature of Skinner's *Verbal Behavior* (1957). He points out, correctly, that I didn't join in the funeral festivities for behaviorism. (I vividly remember a funny talk at an early cognitive science conference in Minneapolis where one of the psychologists—I wish I could remember his name—gave a talk modeled on the forced public "confessions" of faltering Maoists or Castroites: he admitted that he had "committed acts of behaviorism" in his errant youth. He was lampooning the anti-behaviorist fervor of most of the audience, whether or not they realized it—and many didn't think it was funny at all.) Ross then laments my "Skinner Skinned" (1981a) as backsliding. I protest: a close look at that paper, which I still endorse, shows that it is explicitly an attempt to "avoid the familiar brawl and do something diagnostic" (p. 54), exposing Skinner's combination of doctrinaire overstatement and ineffective waffling, and isolating his—and Quine's—distaste for intentional idioms. It should be borne in mind that the next chapter in *Brainstorms*, where "Skinner Skinned" was published, is "Why the Law of Effect Will Not Go Away." I have earned my badge as a circumspect friend of behaviorism, as many philosophers of mind who still don't get it gleefully insist. (For them, "behaviorist" is a term of abuse. We are well rid of the crude doctrine Jack Vickers once called "barefoot behaviorism," but what these philosophers don't realize, apparently, is that their headlong dash away from all hints of behaviorism lands them in the Nagel-Chalmers cul-de-sac.)

I will drag my feet on one point here. Ross is less than satisfied with my proposed role for philosophers in "negotiating the traffic back and forth" between the manifest

and the scientific image. My sense of the point of this project is informed by Sellars' famous definition of philosophy:

> The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term. (1963, p. 1)

Ross says:

> whatever services reconciliation of the manifest and scientific images might render for political and economic support of science, it tends to interfere with the epistemic progress of science. It has this effect because it encourages proliferation of analogies between scientific and folk ontologies, which invariably 'domesticate' the former in the sense of blunting their most radical implications for further conceptual revisions that in turn open roads to new experiments and new mathematical and statistical tools. (p. 41)

Maybe so, on occasion, but I think Ross overestimates the scientists here; they have neither the philosophical prowess nor the invulnerability to massive confusion that this laissez-faire policy presumes. The exercise of finding explanatory paths between scientific and folk ontologies is not always just for the enlightenment of lay audiences; sometimes it is an important reality check (and I use the idiom advisedly) for the scientists and their schemes. We're all in Neurath's boat together, and we have to keep it floating.

2

The main virtue of **Felipe De Brigard's** essay, it seems to me, is a feature it shares with Wilfrid Sellars at his best—not so much the published Sellars as the lecturing Sellars. Wilfrid was a virtuoso blackboard artist, diagramming the logical geography of all the different available positions and showing how, if you adopted his perspective, you could see more clearly the bearing of, say, van Fraassen's constructive empiricism, and Millikan's work, and Jackendoff's, and Churchland's, and Azzouni's and… on the travails of the propositional attitude task force and my problematic part in it. I've learned a lot about the various battlefields I have traversed—and in some regards anticipated and steered clear of—from De Brigard's synoptic but detail-rich account.

> After all, folk psychology is just another theory – unrefined if you want, and operational over a slightly different domain than scientific psychology – but a theory none-the-less. (p. 65)

I wish De Brigard hadn't said that. Or rather, I wish he'd also highlighted the prospect, long urged by me, that part of the continuing confusion on these topics might be the still ubiquitous and unreflective practice of thinking of folk psychology as a *theory*—an assumption perfectly articulated in this sentence. It is this tempting assumption ("What else could it be?") that enables the whole Churchland vs. Fodor dispute so aptly diagnosed here, as well as the theory-theory vs. simulation theory literature, and much else. Sellars' (1956) Jones, the mythical deviser of a theory of mental intervening variables, probably put the theory-theory idea on the stage, but it was I who introduced the term, "folk psychology" in "Three Kinds of Intentional Psychology" (1981b), and from the outset I was on the warpath against treating it as just like a scientific theory. (Well, then, what is it? It's a strategy of interpretation.)

De Brigard shows how the passages he quotes from *C&C* prefigure that campaign; I specifically warned against jumping to ontological conclusions before we tackle the scientific questions, and this caution has been a theme of mine—most effectively, I think, in "Real Patterns" (1991c)—ever since.

In *C&C*, in Chap. 2, "Intentionality," I rather unconvincingly slid by the problems that De Brigard exposes. At that time, I had been persuaded by the opening gambits of Quine and Chisholm that theoretical clarity could be achieved by re-expressing casual mentalistic talk in propositional-attitude formulae. As the years rolled by, it became more and more obvious to me that this was a forlorn quest. Steve Stich and I, on Fulbrights together to Bristol in 1978, decided to write a co-authored paper diagnosing the systematic pathologies of the propositional attitude bandwagon, then in full swing, but we couldn't agree on where to take this shared conviction, so we went our separate ways. I gave it my best shot by writing "Beyond Belief" (1982), by far the hardest philosophical task I have ever completed, and he wrote a whole book, *From Folk Psychology to Cognitive Science: The Case Against Belief* (1983). I considered myself to have paid my dues and shifted my attention to other topics. (Pat Churchland encouraged this decision by dubbing my long, long essay, "Beyond Belief and Past Caring.") Not surprisingly, in retrospect, the propositional attitude task force shrugged off both Stich's diagnosis and mine, and turned back to their research program, which apparently persists to this day. Every decade or so I return briefly to the literature to see if any progress has been made, and De Brigard's critique of the current state of play convinces me, yet again, that they are still spinning the same old wheels. Of course it may be that they saw back in 1981 that my objections were so off target that there was no call to refute them, but over 30 years with no consensus results makes me suspect that they just didn't want to stop playing the game they had mastered.

The dismantling by De Brigard of Paul Churchland's arguments against folk psychology has a few innovations worth noting.

> Blaming the entire apparatus of folk psychology on the basis of just one failure seems a bit exaggerated. For one, I can provide an explanation of the failure in terms of the very same theory: if you hadn't *forgotten* the date, my prediction would have worked just fine. Secondly, it is true that similar extrapolations have proved successful in the past (last Wednesday—remember? —you did actually make it to our appointment). (p. 55)

As De Brigard goes on to note, folk psychology has the resources to identify what sorts of interventions would make forgetting more or less likely. Even if we concede for the moment that folk psychology is enough like a theory to be called a theory, there is nothing circular or vacuous about a theory being able to explain and predict the circumstances under which its predictions tend to be false. *In practice,* in the actual time-pressured world of likelihoods, we not only tolerate but welcome probabilistic predictions (of the weather, of the effects of medicine on our bodies, of weekend traffic jams, and many other things that matter). In such cases, it goes without saying that had the theorist gathered much more data a more accurate prediction could have been made. Ceteris paribus clauses abound, as De Brigard notes, and we don't in general disparage theories that rely on them. I don't recall this point being made before.

Folk psychology is necessary, as De Brigard notes, for devising and executing research programs—as I show in detail in my account of heterophenomenology (1987a, 1991, and elsewhere). In *C&C* I didn't feel the need (in the decadent, waning years of ordinary language philosophy) to support the claim that we can't do without intentional idioms (and the personal/sub-personal distinction—see Frankish, Wilkinson and Roth and my comments on them). Several waves of subsequent eliminativism have tried to persuade us that we are well rid of them, but De Brigard show what a draconian program that would be.

3

Keith Frankish looks at the chapters of *C&C* on thinking and reasoning, and finds previews there of the current enthusiasm in psychology for dual-process theories of thinking, now a major research industry in psychology, most recently made famous outside the academy by Danny Kahneman's (2011) distinction between Type 1 and Type 2 thinking.

I appreciate the restraint of Frankish's constructive and sympathetic account. He doesn't claim that, if only I'd switched my awareness subscripts, it would be obvious that I was the father of dual process theory. As he notes, dual-process theories have been independently invented over and over, a Good Trick with a large basin of attraction, and my particular version was underdeveloped, a philosopher's semi-informed surmise rather than a specifically worked out theory, but at least my reflections were carrying me in the right direction, which is more than can be said for many rival gestures in the direction of theory from other philosophers. What is much more interesting than any priority claim is whether my work, in *C&C* and later, has anything specific to offer to current theory, and here Frankish points to my sketch of an account of how Type 2 thinking gets installed in human brains. That is a big question today, and I do have an inchoate account—the outlines of which Frankish discerns nicely—and am currently working on a considerably more ambitious, detailed, empirically supported, theory.

There is a big problem, which psychologists have been uneven in discerning and more uneven in tackling:

> Type 2 processing seems capable of some prodigious intellectual feats. Indeed, it seems to occupy the role of something rather like a central executive, which can override instinctive, associative, and emotional responses with rational thoughts and decisions. Now, the positing of such an executive system is, of course, a move which Dennett opposes, as being both unexplanatory and neurologically implausible – a central theme of *Consciousness Explained* (Dennett 1991a). (p. 76)

How can you account for the powers of Type 2 thinking without installing an ominously clever *res cogitans* to do the symbol-manipulating? My answer, in short: by recognizing that Type 2 thinking is a *learned, culturally borne, personal-level activity.* As Frankish notes, another problem facing dual-process theorists is to explain how it evolved so swiftly, and, perhaps, only in *H. sapiens.* And yet another is to explain its relationship to Type 1 thinking. How does Type 2 thinking exploit the resources of the mammalian (or more particularly, primate) brain? Frankish presents his answers to these questions, with the help of his interpretation of my own views. This is the best kind of value-added criticism, and I agree right down the

line with his suggestions, and will just highlight a few points here that strike me as deserving extra emphasis.

> Dennett's main concern in this chapter is […] to argue that talk of thinking or reasoning is often simply an idealized intentional characterization of sub-personal information processing operations of which we have no conscious awareness. (p. 79)

Here's one more way of thinking about it: don't make the mistake of treating Type 2 thinking as providing a *process model* for Type 1 thinking—for the same reason we shouldn't view the Type 2 thinking by engineer/designers as the process model for the processes of natural selection! (See my commentary on Dub) Type 1 thinking is fast, parallel, etc., etc. and very good at homing in on excellent results—it gets clever animals through their challenging lives with grace and reliability. When we adopt the intentional stance towards animals in order to explain and predict their behavior, we often treat them as if they were Type 2 thinkers, but that is a crutch for the imagination that should not be seen as committing us to a process model of their Type 1 thinking. Type 2 thinking is a recent add-on, dependent on language, mainly because talking to others and talking to yourself are personal-level actions that play essential roles in installing Type 2 thinking in each of us. Much of Type 2 thinking is in fact talking to yourself—in your native language, not in Mentalese or a "language of thought"—but not all of it is. There is wordless, imagistic (auditory, proprioceptive, tactile, not just visual) exploration that is accomplished by a sort of auto-Socratic method: posing "questions" and seeing what your Type 1 resources can come up with for responses. It is a kind of self-stimulation, in other words, that becomes as "second-nature" as, well, talking to others, some of which is deeply purposive and monitored and some of which is just idle yakking, as noted in Mose Allison's wonderful song, "Your mind is on vacation but your mouth is working overtime." Our internal personal-level activity (cf. Ryle's attempt to answer the question "What is Rodin's Thinker *doing*?") includes not just the hard work of Type 2 thinking but every less disciplined variety of daydreaming and woolgathering.

The Socratic method is in effect an externalization of the private practice of reflection, an exercise in group reflection. In a wonderful passage in the Theaetetus, Plato draws attention to a deep epistemological problem:

> Socrates: Now consider whether knowledge is a thing you can possess in that way without having it about you, like a man who has caught some wild birds—pigeons or what not—and keeps them in an aviary he has made for them at home. In a sense, of course, we might say he "has" them all the time inasmuch as he possesses them, mightn't we?
>
> Theaetetus: Yes.
>
> Socrates: But in another sense he "has" none of them, though he has got control of them, now that he has made them captive in an enclosure of his own; he can take and have hold of them whenever he likes by catching any bird he chooses, and let them go again; and it is open to him to do that as often as he pleases. [*Theaetetus*, trans. Francis M. Cornford (New York: Macmillan, 1957), 197 C-D]

Plato saw that merely possessing knowledge (like birds in an aviary) is not enough; you must be able to get the right birds in your aviary to come when you call. Techniques of self-stimulation designed (unwittingly) to give you access to

your own (Type 1-embedded *and* Type 2) knowledge are the great innovations of Type 2 thinking, and they are thinking tools that must, in the main, be installed—though the installation process, like the acquisition of one's native tongue, has been made easier over the generations by a Baldwin Effect interaction between fast-evolving cultural items and more slowing evolving genetically transmitted design improvements that enhance our ability to use these tools.

It is not just that installed thinking tools give you access to your own embedded knowledge; their installation is what creates the phenomenon of *access* in the first place, by creating the problem of access—and its solution! Type 1 thinking happens automatically, for better or worse, and whatever tracts in the brain get activated do—or fail to do—the right thing. There is no issue of trying to get the right birds to come when you call. But once one has acquired the habit of auto-Socratic exploration there is always the prospect of learning how better to remind yourself of what turns out, on reflection, to have been important. "Next time, it would help if I accessed what I know about X before I decide."

What could guide and control such a process? Frankish has some ideas about this:

> Similarly, self-directed speech acts might be generated pandemonium-style, without antecedent calculation of their structure or likely effects. It is true that, if they are to count as intentional, self-stimulations must be susceptible to some intentional characterization, but this need not be in terms of desires for specific cognitive and behavioural effects and beliefs about how to achieve them. The motivating states might simply be a desire to solve some problem and the instrumental belief that doing *this* (uttering the words that spring to one's lips) may help.
>
> But could pandemonium processes generate the subtle self-stimulations required to support executive control, abstract problem solving, and hypothetical thinking? Where does the intelligence in these acts come from? (p. 84).

He answers his own question. First, many self-stimulations are not particularly intelligent—"chance associations, whimsy, free-wheeling speculations, and so on"—just the sort of hopeful rubbish a pandemonium process would often generate; second, "Self-generated speech and other imagery may not only stimulate cognitive and affective responses, but also trigger further acts of self-stimulation, shaped by those responses" creating "cycles of self-stimulation" that are themselves creative, and

> Third, self-stimulation may be guided by knowledge imparted by culture. Cultural processes may disseminate, not only the trick of self-stimulation itself, but specific applications of it to particular problems. (p. 84)

I think this third point is the key that unlocks the mystery of the "prodigious" power of human thought: in the same way that genetic evolution by natural selection copies and copies and copies the tiny design improvements discerned in each generation, cultural evolution by natural selection (over the last few hundred thousand years) has bench-tested and approved hundreds or thousands of thinking habits and disseminated them widely, in turn creating a huge selection pressure at the genetic level for brains that are good at installing and using these habits. Type 2 thinking is a product of meme-gene coevolution in much the same way lactose tolerance in adulthood is. Dairying is a culturally transmitted practice that creates

selection pressure for lactose tolerance, and auto-Socratic exploration is a culturally transmitted practice that creates selection pressure for brain structures and dispositions that can make the most of these habits.

Once personal-level Type 2 thinking has established itself as the prevailing activity among human beings, further cultural evolution can create whole new phenomena, unknown in the animal world, to exploit these habits. Humor is one of the most distinctive, and it depends on controlling the timing of conscious access (Hurley et al. 2011); a punch line telegraphed loses its punch.

4

Richard Dub baffled me at first, in a very useful way. How could he know the relevant literature so well (not just my work, but Davidson, Lewis, Quine, Ryle, Stich, …) and still not "get it"? Since he lays out the issues better than anyone I've ever encountered, I conclude that I am probably the one who is missing something. What? I'm resisting the temptation to just 'say it again, louder,' and have been casting about for a new way to make my points.

My first hunch is that Dub has underestimated how radical my claim about the rationality assumption is—perhaps out of misplaced charity since he alludes to the incredulity that some philosophers have expressed. His distinction between *individual ascription* and *scientific ascription* allows him to contrast the time-pressured quick-and-dirty attributions of folk psychology with the measured, theoretical posits of scientific psychology, and this allows him to turn my constraint-on-attribution into an empirical discovery. This, he thinks, saves the best features of the intentional stance minus the incredible rationality *constraint*.

> If we interpret agents as rational because we are led to do so by scientific norms of predictiveness, systematization, and empirical adequacy, then rationality need not be a *constraint* on interpretation, nor need it play any sort of role on the *input* side of psychological theory-building. It could be an *outcome*, or *finding*, of (current) psychology that agents are (largely) rational. (p. 104–105)

I, too, want to provide room for empirical discoveries about just how rational we are, and for scientific theorizing about the sub-personal neural mechanisms that subserve (a usefully vague term of art) the phenomena of cognition by successful agents. But I want to demonstrate that the *power* of folk psychology is due to its daring idealization, and that the bold extension of this folk-psychological power to other domains (computer chess programs, Martians, the R&D of natural selection,…) *works* precisely because of its presumption of rationality –and not because, say, the underlying processes of natural selection strongly resemble the processes that occur in believers' brains.

Perhaps a little dramatization can bring this out.

> Curious biologist: I'm baffled by the apparent extravagance of the design of this macromolecule I find in abundance in every bacterium I investigate. I'd like to explain why it has the properties it has.
>
> DCD: My advice to you is to ask *what reason* Nature could have had for devising and protecting such an expensive bit of machinery.
>
> Curious: Are you suggesting I treat natural selection as if it were a rational agent designing the innards of bacteria?

> DCD: Exactly. As Francis Crick often said ("Orgel's Second Law"), evolution is cleverer than you are, so exercise your imagination and remember that Nature is both thrifty and profligate, always willing to settle for a cheap, imperfect "solution" to a design problem, but also willing to throw preposterous resources into the fray (think of the billions of sperm that are wasted every day). And remember: although evolution is remarkably discerning, finding tiny advantages like needles in haystacks and amplifying them, no foresight allowed!

Biologists don't need this advice; they already do this every day. They try (usually unsuccessfully) to refrain from using mentalistic terms in their sober research articles, but they nevertheless use the intentional stance as an imagination-prosthesis to *generate hypotheses to test,* and sure enough, they discover again and again that evolution is a quite reliably brilliant designer of organisms. They aren't trying to prove Orgel's Second Law; they are trying to discover the rationale of the designs of the devices in nature, the better to generate still more hypotheses to test about what how and why these things work the way they do.

Is this just a trick? It's a good trick, an extension of the design stance *that was always latent there.* Remember that using the design stance involves taking on the simplifying assumption that the parts will work as advertised, that they are *good* springs and cogs and axles and bearings. The question that almost goes without saying in every such inquiry is "what could *this* bit be *good* for?" In the case of artifact hermeneutics, there is almost always a (good) reason, moreover, why the parts are arranged as they are, because designers are intelligent, that is to say, rational, and the same holds in the case of evolved entities because natural selection is that good. Of course we don't need a "psychological theory" of natural selection; we already understand the underlying Darwinian algorithms that do all the work. We are just looking for a reasonable rationale for the work that they have done in this case.

The same 'panglossian' idealization works well in folk psychology because people, and animals, and in some regards plants and even bacteria have been well-designed to protect themselves and further their interests. Folk psychology permits folk to make highly reliable predictions with just about zero knowledge of the underlying cognitive mechanisms. (As I have said, the intentional stance taken by itself is *vacuous* as a psychological theory; it presupposes only that the machinery in our heads is well-designed.) Over the millennia we folk have used informal introspection and "intuition" to develop a rather fanciful mythology about what the inner machinery is—desires duking it out, beliefs generated by perceptions piling up in the belief box, images being constructed and perused, intentions being endorsed, urges being suppressed—and some of these folk categories may prove to carve some of neuroscientific nature at the joints, but the utility of folk psychology as a portable sense-maker and hypothesis generator provides scant evidence for this hope, especially given the utility of the same strategy in evolutionary biology (and chess playing computers) where we already *know* that the processes behind the actions we are predicting/explaining are not much, hardly at all, like brain processes—except for the fact that they extract information and put it to adaptive use.

Dub gets close to this with his discussion of electrons. Notice that he doesn't say that we discovered that electrons had negative charge. He said that we discovered subatomic particles with negative charge (and we call them electrons, identifiable or

distinguishable by their negative charge). Similarly, I am claiming, we discovered that intelligent animals have lots and lots of cohering information that they use to guide (appropriately) their actions. We call this information beliefs. We wouldn't call states of a person beliefs that didn't have this delightful property. Electrons are "by definition" negatively charged and beliefs are "by definition" rationally maintained.

Dub supposes we might invent the "schmintentional stance" which found a different way of systematizing the data and deriving reliable predictions. Of course this is possible, and I view cognitive science as engaged in just such an enterprise, finding new categories and states undreamt of in folk psychology. And whenever we encounter likely candidates for such theoretical innovations, we will have to confront the diplomatic/pedagogical (as opposed to metaphysical) question of whether to *identify* these items as none other than the beliefs and desires of folk psychology, or to claim that these items *replace* those obsolete categories. I join Dub in applauding the innovative proposals of Gendler, Schwitzgebel, Egan and Frankish, and don't see my view on the intentional stance as an impediment, aprioristic or otherwise, to such explorations.

I suspect that Dub has some lingering allegiance to the popular idea that we *know* that there are beliefs and are just trying to find the right theory of *them,* while I have been proposing that we consider belief-talk to be a strategy that works well, in spite of all the noisy 'counter-examples' around the edges. As Dub shows, both of my mentors, Quine and Ryle, have contributed to my confidence that this is the wise way to proceed, and I speculate that the residue of disagreement and/or misunderstanding between Dub and me is largely due to his not entirely sharing my enthusiasm for their insights.

Three relatively minor corrections for the record: I think Dub misinterprets Cherniak: it is the believer's finitary predicament, not the attributor's, that leads to the minimal rationality constraint. And Dub misses what I was trying to say about Quine's insouciance about projectionist and normative approaches: Quine noted that even the most ardent projectionist cleans up and normalizes the projections made from his own case, so on any realistic view there is not much room for the different approaches to yield different attributions. Dub is also mistaken when he says that "opinions were introduced in order to *preserve* rationality." Not so. They were introduced to distinguish language-incorporating cognitive states—bets on the truth of sentences, in my sketchy formulation—from other information-bearing cognitive states such as the beliefs (if that is what they are) of animals. (See Frankish, this volume, and my comments on it.)

5

Sam Wilkinson's essay exposes some questions I should have answered long ago, so with the help of his insights, I will try to answer them now: What is the relation between personal level *intelligibility* and *predictability?* What is the relationship between the personal level and the "free-floating rationales" of behaviors and structures of living things? Wilkinson shows me that there is a deeper connection between my early thinking and my more recent thinking than I had realized, and this will permit me to recast some points in what I hope are more persuasive terms. He has done an excellent job articulating my thinking in 1969 about the personal/

subpersonal distinction, and he is right that I drifted away from (but didn't explicitly abandon) two central, Rylean parts of my original claim once I focused my attention on the three stances: I stopped stressing that personal level explanations were non-mechanistic, and favored predictability over intelligibility. I don't know if I ever clearly understood that—and why—I was shifting my emphases in these ways, but I do now, thanks to Wilkinson.

Yes, the personal level is non-mechanistic, and none the worse for that: it is this feature that makes it not just compatible with but congenial with whatever ultimately emerges as the correct mechanistic explanations of the phenomena at the subpersonal level. (On this, see also my dialogue with the evolutionary biologist in my commentary on Dub.) As I would put it today (see Dennett 2014, "The Evolution of Reasons"), it supplements the (mechanistic) answers to the "how come?" questions with the (non-mechanistic) answers to the "what for?" questions. We will still need the personal level because of its role in anchoring intelligibility. Wilkinson also raises an interesting question: the intentional stance seems to render intelligible only a subset of the events that we address at the personal level. When someone says "Ouch!" in response to being kicked in the shin, we understand the meaning of this reaction, but it isn't that the action is shown to be the *rational thing to do, given the subject's beliefs and desires.* Nor is it that we understand this as a merely causal, mechanistic outcome, like the table leg buckling when somebody kicks it. How can I reconcile the *intelligibility* of these cases with the rationality-presupposition of the intentional stance?

First, let me address the shift from intelligibility to predictability. In retrospect I can see that the rationale (possibly free-floating) was this: "intelligibility" has the flavor of Ordinary Language Anti-science Conservatism, an attitude I found insufferable and obtuse ("You scientists go have your fun with mechanistic accounts of bits and pieces of things; we Ordinary Language Philosophers are engaged in appreciating the meanings of acts and ideas, an utterly distinct world off-limits to science.") I wanted to *bridge* the chasm between meaning and mechanism, not *defend* it, and one key element of the bridge was the requirement that intelligibility must have some practical effects, some payoff, some leverage. If rendering some stretch of human activity *intelligible* didn't help us see what to do next, it was just some sort of pointless decoration we couldn't help but indulge in. My motto could have been No Intelligibility without Predictability. I wanted to demonstrate that the intentional stance could do things that the physical stance couldn't do (practically), beating the physical stance at its own game of prediction. *That's* why the personal stance is ineliminable; not for Wittgensteinian "reasons" (explanations have to stop somewhere) or Strawsonian "reasons" (we just *are* the sort of creatures who harbor resentment), but for an ultimately biological reason: the personal level, by making life somewhat predictable, helps us live safer, easier, more productive lives. If I'd made that point, it would perhaps have forestalled the common objection to the intentional stance along the lines of "but we're not interested in *predicting* our companion's every move; we're interested in *understanding* it." To which my reply is: you may not appreciate that you are engaged in prediction, but you are, automatically and involuntarily, and there would be no understanding without it. I have often said that the job of the brain is to "produce future," a claim that is becoming more and more obvious as Bayesian approaches to cognition come to be appreciated.

Of course I also wanted to stress the continuity between our personal level attributions and the attributions of computer programmers, biologists, and other scientists working with designed systems. So I would decline Wilkinson's suggestion that "there is nothing metaphorical" about attributing a belief to a person, in contrast with attributing a belief to a chess-playing computer—if that means there is a sharp line between the two practices. I see them, and have always seen them, as on a continuum, with *more* "literal" attributions at one end, and highly fanciful (but still explanatory and hence justifiable) attributions at the other.

Now what about the intelligibility of saying "Ouch" when in pain? *Why* do we shudder, wince, tremble, smirk, sigh, flinch, scream, groan,…? Emotional reactions are not intentional actions, but they do typically get explained with ineliminable appeal to the intentional stance, because they are often the involuntary (and typical) responses to beliefs and desires. In fact, some of the most secure clues to intentional stance attributions are involuntary emotional responses. The children shrieking with delight at the puppet show because they believe Punch believes that Judy is in the box is blue-chip evidence that they are capable of attributing false beliefs, no matter what they can or can't say, intentionally, in response to adult's questions. (See Dennett, "Beliefs about beliefs" 1978, and the Sally-Ann industry—false-beliefs-tasks—that arose from it.)

Some emotional reactions shade seamlessly into voluntary, deliberate, intentional responses to the same circumstances. Ducking intentionally is continuous with flinching; shuddering and trembling stand in between fainting or collapsing in despair and fleeing (intentionally) from something feared/believed harmful, etc. They are *intelligible* responses in part because they are *familiar* symptoms of *typical* beliefs and desires, which we effortlessly learn to rely on (unless we are autistic, for example) in our largely involuntary adoption of the intentional stance. A child who has not yet witnessed blushing embarrassment (or fury) may find the first few bright red faces unintelligible, but will soon catch on. But they are also often intelligible because in spite of not being intentional actions, they have free-floating rationales that we may vaguely appreciate (and sometimes we're wrong about them).

The backbone of personal level intelligibility is the rationality assumption, without which body language and facial expression would be almost powerless as clues. It would be an interesting (and fun) experiment to take a film—a romantic comedy, let's say—and re-edit it with the help of some computer graphics so that actions and facial expressions that made effortless sense in context were now utterly baffling because the intentional stance could get no purchase on what the characters were engaged in trying to do.[1] I might go so far as to say that intelligibility just *is* predict-

[1] Woody Allen's first film, *What's Up Tiger Lily?* (1966) was the inspiration for this suggestion, but I'm proposing something different, and more radical. Allen bought the rights to a Japanese James-Bond-type film, threw away the dialogue and the plot and dubbed it with an entirely different story, dealing with a lost recipe for egg salad. It's goofy and Dadaist and fun, but I'm imagining leaving the dialogue intact, but rearranging brief scenes in such a way as to make them unintelligible in spite of the fact that they were normal bodily responses to very particular circumstances. I predict the results would be striking and unsettling.

ability from the intentional stance, but it is important to recognize that the rationality assumption of the intentional stance also has work to do quite independently of the personal level. (See also my comments on Dub.) Many human and animal behaviors have free-floating rationales (Dennett 1983, 2013, 2014) that are *not* personal level explanations, though they are often sloppily described as if they were. When gazelles *stott* (make those amazing leaps while being chased by lions) they are signaling that they are healthier than average and hence harder to catch, and the lions "believe" them, and turn their attention to other gazelles that can't stott. Neither the gazelles nor the lions need to understand these signals, but this is the free-floating rationale for this otherwise baffling behavior, well confirmed by both evidence and theory. We use the intentional stance to render the behavior intelligible (and it is manifestly *not* a mechanistic explanation, since no agent—no gazelle, no lion, no intelligent designer—formulates the rationale in anything like a language of thought (or Language of Divine Thought). When zoologists speak, loosely, of the gazelles signaling, it looks for all the world like a personal level attribution (cf. "She's signaling to you that it's time to leave the party.") and it may give some romantics the impression that the gazelles are being attributed great wiliness and appreciation of lion-psychology. (Those clever gazelles! They appreciate how to bluff the lions into leaving them alone!). And when evolutionary psychologists speak, loosely, of women being "coy" because they have a greater "investment" in reproduction than men (in time—9 months—and precious eggs), this is the free-floating rationale for an undeniable asymmetry in the animal (and human) world but not at all a personal level explanation! Evolutionary psychologists are *not* claiming that women (in general or in particular) have a miserly attitude towards their precious ova and are disingenuously assaying men for their genetic fitness with every ploy. Failure to appreciate this is probably a major source of the otherwise bizarrely overwrought negative reaction to evolutionary psychology by many deploring critics.

A personal level explanation is one that a person can acknowledge, report, appreciate, evaluate (or, of course, dissemble about—but you have to be aware of it to dissemble about it). In the case of the free-floating rationale for a preference, habit, tendency, or reflex reaction, for instance, it counts for nothing when a person claims not to have considered, or to understand, or to accept, it. It may still be the (non-mechanistic, rationality-presupposing) intentional explanation of the sub-personal arrangement that provides the *how come* explanation. Personal level explanations of various human features are notorious for bottoming out rather suddenly in vacuity, as in "We love jokes because they are so funny!" or "I like sweet things because they taste so nice!" "Her dance arouses me because it is so sexy!"—all true enough but uninformative as answers to the perfectly legitimate *what-for* questions that remain untouched (Hurley et al. 2011).

The line between personal level explanations and free-floating rationales is porous. In the case of non-human animals, it is particularly easy to see that a personal level attribution of belief or desire may seriously exaggerate the presumed understanding of the animal while still speaking truly of the underlying rationale of the behavior. That the dog wants to catch the squirrel and believes the squirrel is still

in the tree it is standing under is plausibly "personal level" for the dog, but the dog's belief that strangers are not to be trusted or desire for fatty acids in its diet can be seen as more properly free-floating rationales for tendencies of which the dog has scant—if any—understanding. In the case of human beings, I think that much human behavior we generally treat as rationally intended and well understood is at best only dimly (or retrospectively) understood. If you have ever made a move in a chess game the brilliance of which only later dawns on you, but claimed that you had the insight all along, you know how easy it is to fool others—and even yourself—about how intelligible *in prospect* your own behavior is to yourself. We still consider such a chess move a personal level action (unless a piece is accidentally nudged, say) because it occurs in a paradigmatic setting of rational agents in competition. Similarly, we treat "impulse purchases" and other responses to covert manipulations as personal level phenomena even when people are demonstrably confused or ignorant of the influences on their choices because these are transactions between "consenting adults" who are presumed to be rationally guiding their actions.[2]

When we look at the personal/subpersonal level distinction in the context of mental illness, as Wilkinson shows, we find a rather different porous boundary, created by the (ultimately mechanistic) pathology in the subpersonal systems of perception and control, leaving many attributions problematic at the personal level. Here I think Wilkinson slightly misstates the case to be made when he says that top-down and bottom-up theories, while both making the personal/subpersonal distinction, differ "about substantive, empirical facts about what is going on inside these patients." (p. 124) They may not differ on these (largely still unknown) facts, but only on what the threshold of understanding for personal level attributions should be. As Wilkinson points out, delusional patients typically fail to act on their delusional claims in ways that would tend to rescind the attribution of belief were it not for the sincere avowals by the sufferers. I also disagree with his claim that "when the chess computer malfunctions, it is so different from us that we would never ask to render its malfunctioning behavior *intelligible*, let alone expect to be able to do so." We wouldn't expect the computer to do so. (Do chess playing computers have a personal level? Not yet, I would say.) But we *do* often render their malfunctions intelligible by using the intentional stance. One of my favorite real life examples, often cited by me, was Rich Greenblatt's casual observation of a rival chess playing program that "it thinks it should get its queen out early." In a single stroke this comment rendered a great deal of that program's behavior intelligible, but I guess it should count as a free-floating rationale. (And note: even the intelligent designer of that program didn't contemplate or consider that proposition or attribution in the course of designing the program.)

[2] Felipe De Brigard, editing this passage, made a useful comment that deserves to be quoted: "Reasonable behavior need not be behavior that responds to reasons, or that is brought about in response to reasons. I like it. Many traditional philosophers will disagree, of course. To them I'd say, in the guise of Don Quixote, 'let the dogs bark, Sancho. It is a sign we are on track.'"

6

Martin Roth repairs "a conspicuous absence" in the current controversy over embodied cognition: the honoring of the personal/sub-personal distinction. I think he is right about its application, and in retrospect I am as surprised as he is that the combatants have ignored it, especially since both Andy Clark and Fred Adams have been interacting with me for decades on other topics. Unaccountably I never thought to propose my own distinction to them, but I expect Andy will endorse Roth's friendly amendment. It will be interesting to see if Adams and Aizawa have a response.

Roth's reconstruction of my disagreement with Fodor over Ryle is right on target. I wish I'd seen then as clearly as I do now, thanks to Roth's analysis, just what the core of our disagreement was. When I wrote *C&C,* Fodor was known to me only through his papers on meaning and linguistics with Jerry Katz, which were all the rage in Oxford and elsewhere, one of the opening salvos in the siege that pretty well extinguished ordinary language philosophy, but I didn't see myself as having a dog in that particular fight. (I tended to side with Fodor and Katz, as part of my growing interest in bringing science to bear on philosophical issues.) That explains why there are no references to Fodor in *C&C*. After I had sent my manuscript to Routledge & Kegan Paul, my colleagues at Irvine, Joe Lambert, Gordon Brittan, and Jack Vickers proposed a discussion group on Fodor's new book, *Psychological Explanation* (1968)*,* which provoked us all, in different directions. I remember that we terrorized the graduate students who sat in on it with the vehemence of our attacks on each other's interpretations and arguments. We were all dear friends, but a mark of that friendship was our enthusiasm for blasting away at each other with abandon—offering hooting *reductios*, sarcastic "parody of reasoning" putdowns and all manner of scoffing and name-calling—philosophical debate comes to the locker room. The ideas in "Intentional Systems" (1971) were almost literally hammered out in that discussion group, so that was the first time I got lifted by Jerry, the human trampoline. As I have said before, if I can see farther than others it is because I've been jumping on Jerry.

What strikes me on reading Roth's essay is how well I'd anticipated Fodor's subsequent (mis-)reading of Ryle in my warnings. I met Fodor soon after moving to Tufts in 1971, and we soon got to thrash out the issues in person, on many occasions, in a discussion group of Boston-area philosophers, and on Jerry's sailboat, Insolvent, but that's a tale for another day.

Roth says

> if Adams and Aizawa are correct that "Underived content arises from conditions that do not require the independent or prior existence of other content, representations, or intentional agents" (2010, p. 32), it will turn out that the intentional contents and processes of people – brain-bound *or* extended – are derived. (p. 141)

I agree, but perhaps in a somewhat different sense of "derived" than Adams and Aizawa intend. In my debate with Searle on original and derived intentionality, I point out how, on an evolutionary account of the birth of content, all the content in the nervous systems of organisms turns out to be just as "derived" as a written

shopping list (*The Intentional Stance,* "Evolution, Error and Intentionality" (1987b)). I suspect that for many people in the field, this appears to be a bridge too far, but I think it is the key insight needed to break away from what Quine called the museum myth of meaning. (See my "With a little help from my friends," in Ross and Brook, eds., 2000, and "Radical Translation and a Quinian Crossword Puzzle," in *Intuition Pumps*, 2013). If you try to be a more staunch realist about content than this, you inevitably find yourself drifting down Searle's stream to an ultimately mysterian view of original intentionality.

7

Ellen Fridland's essay clarifies constructively what I said about intelligence and learning, clearing out a few clouds and sharpening the focus. I find nothing substantial to disagree with, but will avail myself of the opportunity to build a few more wrinkles into her account, and, first, correct a factual error of mine that she has innocently propagated.

Both Doug Hofstadter and I were struck by the Wooldridge passage she quotes, and Doug was inspired to coin the very useful term "sphexishness" in honor of these wasps, but:

> We have recently learned that Wooldridge gave us—as popular science writers so often do—an oversimplified sketch of the phenomenon. The psychologist Lars Chittka wrote to me, quoting from the work of Jean-Henri Fabre (1879!), which had apparently been the source for Wooldridge, who, if he had read on in Fabre, would have found that in fact only some Sphex wasps are sphexish! In fact, Fabre was eager to make the point. If at first blush you thought Sphex was clever, and at second blush you thought Sphex was stupid, try third blush, and find that some Sphex are not so sphexish after all. Chittka sent me the German translation of Fabre (I still haven't located the French) which includes the following sentence: "*Nach zwei oder drei Malen… packt ihre Fuehler mit den Kieferzangen und schleift sie in die Hoehle. Wer war nun der Dummkopf?*" ("After two or three times,… she grabbed her [the prey's] antennae with her pincers and slid it into the hole. Now who's the dummy? (Dennett 2013, p. 398)

So now we can all go on using the term "sphexishness" with clear consciences, knowing that it is something of a misnomer, but too well established to abandon.

Fridland rightly highlights the normativity that brings flexibility and manipulability into the picture (since intelligence isn't magic), but these features thereby also frustrate—predictably, I would say—any attempt to capture intelligence inside any fixed definitional fence. For instance, an intelligent agent has the intelligence to adjust its interests, so what is in the 'best interests' of an agent can change almost indefinitely: suicidal projects are not ruled out, for instance, if they further the highest goals of the agent. As I put it in *Breaking the Spell* (2006):

> Whenever an agent—an intentional system, in my terminology— makes a decision about the best course of action, all things considered, we can ask from whose perspective this optimality is being judged. A more or less standard default assumption, at least in the Western world, and especially among economists, is to treat each human agent as a sort of isolated and individualistic locus of wellbeing. What's in it for *me*? Rational self-interest. But although there has to be something in the role of the self—something that answers the *cui bono?* question for the decision-maker under examination— there is no necessity in this default treatment, common as it is. A self-as-ultimate-beneficiary can in principle be

> indefinitely distributed in space and time. I can care for others, or for a larger social structure, for instance. There is nothing that restricts me to a *me* as contrasted to an *us*. I can still take my task to be looking out for Number One while including, under Number One, not just myself, and not just my family, but also Islam, or Oxfam, or the Chicago Bulls! The possibility, opened up by cultural evolution, of installing such novel perspectives in our brains is what gives our species, and only our species, the capacity for moral—and immoral—thinking. (p. 176)

Her definition of learning nicely incorporates the tremendous changes that are wrought in us by cultural inculcation, and she notes that it is when we are what I have called Gregorian creatures that we particularly surpass in intelligence all other learning agents on the planet. What gets in the door by this route are a lot of thinking tools from which we can benefit without entirely understanding. As

Andy Clark (1997) puts it, "We use intelligence to structure our environment so that we can succeed with less intelligence. Our brains make the world smart so we can be dumb in peace!" This can sometimes appear to be "cheating" when we consider (or measure) intelligence. Can you bring your pocket calculator or laptop to the exam? It depends on many factors. Senator Ted Kennedy was intelligent enough to realize that he was not intelligent enough on his own to make good decisions on many issues so he appointed the smartest advisors he could find and listened to them. Now that's smart! But what are the limits?

As we offload more and more of our opportunity-generation-and-assessment chores to handheld electronic thinking tools (or trusted human advisors, for that matter), are we heading into transhuman sphexishness? There is no easy answer to that question, which I have been pondering for decades. It was one of the issues that inspired me, along with my colleague George Smith, to create the Curricular Software Studio at Tufts back in 1985. We had a metaphor: there were two ways of improving human muscle power: the bulldozer way and the Nautilus machine way. The first way lets you move mountains but you may still be a weakling; the second uses technology to build up your personal strength. We set out to create Nautilus machines for the mind, "imagination prostheses" that could enhance your *understanding,* not just give you the right answers. It is possible, and desirable, because—use it or lose it—if you delegate the hard questions to your tools, you'll have no way of knowing if the answers you get are right.

Another point of Fridland's I want to enlarge upon is her observation that an implication of the manipulability requirement is that

> intelligence becomes a personal-level phenomenon. This is because manipulability requires global, integrated, centralized, hierarchical processes that are not available to subpersonal systems. That is, to be manipulated, a state must be targeted by higher-order states or mechanisms. The requirement that intelligent states are personal-level accords nicely with our intuitions about intelligence since, at the very least, the requirement that behaviors, processes, or representations be manipulable puts intelligence in the same realm as, for example, rationality and knowledge. (p. 151)

More pointedly, it is this feature that gives consciousness *real work to do* (if you hold a sane view of consciousness, ignoring zombies and the so-called Hard Problem). Meta-representation is a core strategy of greater intelligence (Fridland

aptly cites Clark and Karmiloff-Smith on this, and for a much more detailed examination of the issue see Stanislas Dehaene's recent book, *Consciousness and the Brain: Deciphering how the Brain Codes our Thoughts* (2014), which does an excellent job of ushering the "consciousness as mere epiphenomenon" view off the stage.) The personal level is constituted by *what persons can share and discuss about what they are doing,* and that practice is, quite obviously, the key to human intelligence. Einstein, forced to grow up alone on a desert island, without language, would be profoundly disabled, cognitively, however "gifted" at birth by his genetic endowment.

Finally, I particularly applaud her footnote 3 on her methodology, which is also mine.

> The notion of intelligence that I am pursuing is a scientific notion. As such, my methodology will not be conceptual analysis. In this kind of endeavor, if various counterintuitive consequences result from my account, these will not immediately count as a *reductio* of the position. After all, science is often counterintuitive. Still, I hope to illustrate that what we think of as intelligence is already, to a large extent, in line with the claims that I am making here. As such, I would like the notions of learning and intelligence that I put forward to correspond to ordinary intuitions as much as possible. However, I do not insist that if ordinary intuitions conflict with the account I am offering, then the account is wrong. On my approach, it may turn out that we have *empirical* or *methodological* reasons that trump our ordinary intuitions. Intuitions ought to be considered, but they ought not to be the final arbiters. (p. 144, fn. 3)

This strikes me as just obvious good sense in the twenty-first century, but I find a surprising number of philosophers who resist it. An example of her methodology at work in this essay is her recognition of the role of higher-order or meta-representational states. I don't think this would emerge from any pure "conceptual analysis" of the concepts of intelligence or learning. Once noted, it is quite intuitive, as she observes, but its warrant arises from empirical work in psychology, neuroscience and related fields, not from armchair reflection.

8

John Michael's proposal of a developmental loop that sustains and refines the intentional stance as a predictive/explanatory strategy usefully builds on the earlier insights of McGeer and Mameli, and I welcome this as an enlargement and improvement of my account of the intentional stance. His survey of the empirical literature (Gergely, Csibra, et al.,…) is right on target, and as he notes, it is all at least consistent with, if not directly supporting, the idea that concept-acquisition or concept-mastery is itself a gradual, approximating phenomenon. And what defines the gradient up which this competence marches? Rationality, in the neutral sense of cognitive competence, whatever that comes to. Here are two apparently very different ways of putting the claim:

> in Bayesian terms: children come to have ever more accurate, reliable, high-fidelity expectations.
>
> in propositional-attitude terms: children see more and more of the implications of the propositions we are boldly attributing to them.

The intentional stance takes "propositional attitudes" from folk psychology as a way of alluding to what is learned, while the Bayesian also takes something like propositions for granted without going into the details (*Expectations*? Just what is an expectation, and how many of them can you distinguish…?) In either case it is important to recognize that propositions are idealizations on their own, as I argued in *C&C* in my example of the child who says, "Daddy is a doctor" (p. 183). (See also my discussion of the sorta operator, in *Intuition Pumps.*) Just what, exactly, does the child believe at time *t*? For convenience we can choose one or another sentence as the best expression of what we are getting at, and plug it into our Bayesian formula, or into our propositional attitude attribution, but except, maybe, for rare artificially sharp-boundaried categories, these are always idealizations.

Notice, by the way, that as soon as we permit ourselves to talk, as Michael does, of infants who "partially master" concepts, we have left Fodor and many of the "propositional task force" operatives on a distant shore. For them, Fregean grasping is all or nothing—you either have the concept HORSE (or the concept SCHMORSE) or you don't. They are taking a bold idealization as if it were a description of brute facts and trying to theorize about it.

Michael's idea about the developmental loop permits us to put the two ways of thinking about rationality and intentionality together: children start Bayesian, like all young animals, and are gradually intellectualized by language, so that propositional-attitude talk, always idealized, begins to do more and more justice to their psychological states. They come to have opinions, but these bets on the truth of sentences are themselves a gradually blooming and refining matter, as the child's opinion expressed as "Daddy is a doctor" reveals.[3]

The upshot of Michael's developmental loop is his suggestion that as it recursively feeds on its own outcomes,

> this entanglement of pattern recognition, on the one hand, and pattern etiology on the other, provides an additional justification for the belief that those patterns indeed exist, because our recognition of the patterns enables us to further embed them in their respective target systems. (p. 177)

Yes, indeed, but this passage raises the ominous specter of a community-wide delusion that is innocently supported by the new initiates. This probably has happened. Perhaps belief in witches is like this. When everybody has the category, and knows what the defining marks of witches are, group consensus is achievable about not just the existence of witches as a general proposition but also about the identification of particular people as witches. If you've been raised to look out for witches, you'll soon be pretty good—consilient with your peers—at the task. Might the intentional stance be nothing more than a persistent communal delusion then? I expect Paul Churchland and other eliminativists would be tempted to accept this

[3] I agree with Michael's suggestion that the "causal realist option put on the table here remains compatible with the general spirit of Dennett's theory," (p. 179) so there is no difficulty maintaining continuity of reference as both attributors and attributes revise their TOMs (though I continue to dislike the use of "theory" in this context).

little gift from me, and declare that I had finally seen the light! But, I reply, there is a striking difference: the patterns discovered and highlighted by the intentional stance are prodigiously predictive of not only those who have been enculturated to adopt the stance, but of animals, well-designed robots and chess-playing computers, and indeed of natural selection, the blind watchmaker. So I must disagree with one final passage:

> A further insight generated by the developmental perspective is that it is perhaps not an assumption of ideal rationality that constitutes the core of the intentional stance as an interpretive strategy but an assumption of culture-specific imperfect rationality. (p. 180)

I am still going to resist this, since I think that culture-specific imperfections are largely elaborations of, rather than alternatives to, ideal rationality. When you learn about witches, your expectations are flavored by witch-categories, but your nervous system is still engaged in optimizing its expectations in these terms.

9

I am grateful to **Pete Mandik** for doing such a good job policing the HOT topic, a task I decided to leave to others in the twenty-first century. (My last of three or four responses to Rosenthal was Dennett 2000. See earlier 1991b, 1993a, b) I see that in the meantime there has been a lively debate, but on Mandik's reading, it seems—to my relief—to be about to land back on my playing field after all. First person operationalism is a not so strange attractor, and it has accumulated not only allies in cognitive neuroscience, but an impressive and growing bounty of experimental results. For the latest, see Dehaene (2013). In *CE*, I made it clear that I thought David Rosenthal's HOT theory was right about *something*—and something important. His categories were, and still are, resolutely folk-psychological—"thoughts" and "beliefs"—but he has been driven, appropriately, to stretch their sense: unconscious thoughts are not just acceptable but required by Rosenthal's HOT theory, and a non-relational reading of higher-orderedness is, well, in order. What he had found was an almost-folk-psychological way of expressing something important about the relation between (human) consciousness and communication: since we can report our conscious experiences, we must have thoughts (occurrent or episodic beliefs, if you prefer) to be *expressed* by those reports. Hence to have a conscious experience is *ipso facto* to have a thought to the effect that you are having it. I endorsed the strong tie to reportability in human consciousness (we'll consider non-human consciousness later), while finding Rosenthal's way of putting it still too Cartesian, depending as it does on an unanalyzed *res cogitans,* the thinker of those thoughts, "at the top."

Once we acknowledge unconscious higher-order thoughts, why should being the object of just any old higher-order thought secure the elevated status of consciousness? Rosenthal must be supposing there is a privileged variety of higher-order thoughts, unconscious but somehow central or dominant, that secure this status. It's like the difference, I suggested, between being famous or influential and being *known by the King* (Dennett 2000). In some countries being known by the King is both sufficient and necessary for being influential. In some more democratic or even

anarchic regimes, there is no King whose cognizance is obligatory for influence. It would be better, I urged, to capture Rosenthal's points in explicitly sub-personal terms: fame in the brain, for a start, which doesn't depend on being known by the Emperor, because there is no Emperor. Ray Jackendoff's (1987, 2011)—and now Jesse Prinz's (2012)—vision of consciousness as an intermediate-level cognitive phenomenon with "higher" but unconscious processes doing much of the interpretive and reactive work, are elaborations of this idea with much to recommend them (see Dennett 2015)

> I have always treasured Voorhees' outrage on this score:
>
> Daniel Dennett is the Devil. […] There is no internal witness, no central recognizer of meaning, and no self other than an abstract 'Center of Narrative Gravity' which is itself nothing but a convenient fiction.… For Dennett, it is not a case of the Emperor having no clothes. It is rather that the clothes have no Emperor. (Voorhees 2000, pp. 55–56)

Of course the clothes have no Emperor—it wouldn't be a theory of consciousness if the Emperor were still there, witnessing and reacting to all the goings on in the Cartesian Theater.

References

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York: Viking Press.

Dennett, D. C. (1971). *Intentional systems* (pp. 87–106). 68: Journal of Philosophy.

Dennett, D. C. (1975). Why the law of effect will not go away. *Journal for the Theory of Social Behaviour, 5*, 169–187. Reprinted in W. Lycan (Ed.). (1990). *Mind and cognition: A reader.* MIT Press.

Dennett, D. C. (1978). "Beliefs about beliefs," (commentary on Premack, et al.). *Behavioral and Brain Sciences, 1*, 568–570.

Dennett, D. C. (1981a). *Brainstorms. Philosophical essays on mind and psychology* (pp. 53–70). Cambridge, MA: MIT Press.

Dennett, D. C. (1981b). *Three kinds of intentional psychology*. In R. Healey (Ed.), Reduction, time and reality (pp. 37–60). Cambridge: Cambridge University Press.

Dennett, D. C. (1982). Beyond belief. In A. Woodfield (Ed.), *Thought and object: Essays on intentionality.* Oxford University Press.; excerpted in (1996). *From beyond belief: Notional attitudes.* In A. Pessin, & S. Goldberg (Eds.), *The twin Earth chronicles: 20 years of reflection on Hilary Putnam's the meaning of meaning* (Chap. 9, pp. 161–179). Armonk: M.E. Sharpe.; reprinted in the *Intentional stance* (pp. 117–202).

Dennett, D. C. (1983). "Intentional systems in cognitive ethology: The 'Panglossian Paradigm' Defended," (with commentaries). *Behavioral and Brain Sciences*, *6*, 343–390. German translation, "Intentionale Systeme in der kognitiven Verhaltensforschung," In Münch, D. (Ed.). (1992). *Kognitionswissenschaft: Grundlagen, Probleme, Perspektiven.* Frankfurt: Suhrkamp. Excerpt reprinted as "Intentionality in primate social cognition." In R. W. Byrne, & A. Whiten (Eds.), *Social expertise and the evolution of intellect.* Oxford University Press. Italian translation published in *Pegaso*, Centro Documentazione Rovigo.

Dennett, D. C. (1987a). *The intentional stance. Cambridge*, MA: MIT Press.

Dennett, D. C. (1987b). Evolution, error and intentionality. In *The intentional stance* (pp. 287–322). Cambridge, MA: MIT Press.

Dennett, D. C. (1991a). *Consciousness explained*. Boston/New York: Little, Brown.
Dennett, D. C. (1991b). "Lovely and suspect qualities," (commentary on Rosenthal, *The independence of consciousness and sensory quality*). In Villanueva, E. (Ed.). (1991). *Consciousness* (SOFIA Conference, Buenos Aires, pp. 37–43). Atascadero: Ridgeview.
Dennett, D. C. (1991c). Real patterns. *Journal of Philosophy*, *88*, 27–51.; reprinted in Lycan, W. G. (Ed.). (1998). *Mind and cognition: An anthology.* Blackwell Publishers.; translated into French and reprinted in *Philosophie de l'esprit: une anthologie*, Librairie J. Vrin Philosophique, eds. Pierre Poirier and Denis Fisette; translated into Spanish and reprinted in "Sabemos como se aprende?" published by the Ministerio de Educacion del Peru, 2001, pp. 201–233; translated into Polish and reprinted in *Analityczna Metafizyka Umysłu*, Warsaw 2008, pp. 299–325.
Dennett, D. C. (1993a). "Living on the edge", (reply to seven essays on *Consciousness explained*). *Inquiry, 36*, 135–159.
Dennett, D. C. (1993b). "The message is: there is no medium" (reply to Jackson, Rosenthal, Shoemaker & Tye). *Philosophy & Phenomenological Research, 53*(4), 889–931.
Dennett, D. C. (2000). "With a Little help from my Friends" response to Rosenthal. In D. Ross, A. Brook, & D. Thompson (Eds.), *Dennett's philosophy, a comprehensive assessment* (pp. 327–388). Cambridge, MA: MIT Press.
Dennett, D. C. (2006). *Breaking the spell*, *religion as a natural phenomenon.* New York: Viking Press.
Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York: W.W. Norton & Company.
Dennett, D. C. (2014). The evolution of reasons. In H. D. Muller & B. Bashour (Eds.), *Contemporary philosophical naturalism and its implications* (pp. 47–62). New York: Routledge.
Dennett, D. C. (2015). The Friar's fringe of consciousness. In I. Toivonen et al. (Eds.), *Structures in the mind: Essays on language, music, and cognition in honor of Ray Jackendoff*. Cambridge, MA: MIT Press. forthcoming.
Fabre, J.-H. (1879). *Etude sur les moeurs des Halictes* [I have no info on the publisher.
Hurley, M., Dennett, D. C., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. Cambridge, MA: MIT Press.
Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.
Jackendoff, R. (2011). *User's guide to thought and meaning*. New York: Barnes and Noble.
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Geroux.
Plato. (1957). *Theaetetus*. Trans. F.M. Cornford. New York: Macmillan, 197 C-D.
Prinz, J. (2012). *The conscious brain: How attention engenders experience*. Oxford: Oxford University Press.
Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *The foundations of science and the concepts of psychology and psychoanalysis* (pp. 253–329). Minneapolis: University of Minnesota Press.
Sellars, W. (1963). *Science, perception and reality*. London: Routledge & Kegan Paul.
Skinner, B. F. (1957). *Verbal behavior*. New York: Copley Publishing.
Stich, S. (1983). *From folk psychology to cognitive science: The case against belief*. Cambridge: MIT Press/A Bradford Book.
Voorhees, B. (2000). Dennett and the deep blue sea. *Journal of Consciousness Studies, 7*(3), 53–69.