A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology

Edited by
Timothy P. Racine
Kathleen L. Slaney

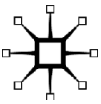
A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology

A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology

Edited by

Timothy P. Racine and Kathleen L. Slaney
Simon Fraser University, Canada

palgrave
macmillan



Softcover reprint of the hardcover 1st edition 2013

First published 2013 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries

ISBN 978-1-349-35031-5 ISBN 978-1-137-38428-7 (eBook)

DOI 10.1007/978-1-137-38428-7

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

Contents

List of Figures

Acknowledgements

We are very grateful to Alan Costall, Hans-Johann Glock, Peter Hacker, Nicole Hausen and Michel ter Hark, Daniel Hutto, Oskari Kuusela, Michael Maraun, Danièle Moyal-Sharrock, John Preston and Severin Schroeder, Diane Proudfoot, Joachim Schulte, Stuart Shanker and Devin Casenhiser, Wes Sharrock and Leonidas Tsilipakos, Michael Tissaw and Meredith Williams for their insightful and inspiring contributions to the present volume. We also wish to thank Melanie Blair, Pri Gibbons and Brendan George at Palgrave Macmillan for their guidance and support through the process and Leona Ferguson for editorial assistance.

Notes on Contributors

Devin M. Casenhiser is Assistant Professor at the Department of Audiology and Speech Pathology at the University of Tennessee, United States. His research explores functional (i.e., communicative) and cognitive factors that affect language learning in neurotypical and autistic populations. He has published research investigating the effects of homonymy on lexical acquisition, how the distribution of lexical input affects children's ability to learn grammatical constructions, and together with colleagues in Canada, he has developed and tested a social-interaction-based intervention designed to improve the social-communication abilities of children diagnosed with autism spectrum disorders.

Alan Costall is Professor of Theoretical Psychology at the University of Portsmouth, UK. In his theoretical and historical work, he has been examining the origins of the dualistic thinking that pervades modern psychology: mind vs. world, mind vs. body, subject vs. object, individual vs. society, biology v. culture. His work has been an attempt to develop an alternative approach to mainstream cognitivist psychology, based on the mutuality of animals and environments, and people and their situations. His recent publications include *Against Theory of Mind* (2009) and a 2012 target article in *Behavioural & Brain Sciences* advocating for a 'second person neuroscience.'

Hans-Johann Glock is Professor at the Institute of Philosophy at the University of Zürich, Switzerland. His research interests include the philosophy of mind, the philosophy of language, the history of analytic philosophy, and Wittgenstein. His current work concerns concepts, the normativity of language, and animal minds. His recent publications include *Wittgenstein: A Critical Reader* (2001), *Quine and Davidson on Language, Thought and Reality* (2003), *What Is Analytic Philosophy*? (2008), and *Wittgenstein and Analytic Philosophy* (2009, edited with John Hyman).

P. M. S. Hacker is Emeritus Research Fellow in the Department of Philosophy at St John's College, Oxford, UK, and is a leading authority on Wittgenstein. He is author of the four-volume *Analytic Commentary on the* Philosophical Investigations, the first two volumes of which were

co-authored with G. P. Baker, and in 2009 he translated and edited the fourth edition of *Philosophical Investigations* with J. Schulte. He has also written extensively on philosophy and the neurosciences, most recently *Philosophical Foundations of Neuroscience* (2003) and *History of Cognitive Neuroscience* (2008), co-authored with M. R. Bennett. Finally, in 2007, he published the first volume of what is intended to be a three-volume series on human nature, *Human Nature: The Categorial Framework*.

Nicole Hausen is a doctoral student in the Graduate School of Philosophy at the University of Groningen, the Netherlands. Her research focuses on the later writings of Ludwig Wittgenstein. The aim of her Ph.D. project is to provide a comprehensive, conceptual interpretation of Wittgenstein's remarks on emotion and to show how Wittgenstein's account of emotion fits within the broader context of his philosophy of psychology. One major relation being explored, for example, is the connection between emotion and aspect seeing which is pursued in her contribution to the present volume.

Daniel D. Hutto is Professor of Philosophical Psychology at the University of Hertfordshire, UK. His research is a sustained attempt to understand human nature in a way which respects natural science but which nevertheless rejects the impersonal metaphysics of contemporary naturalism. His recent projects have focused on consciousness, intentionality and everyday social understanding. His recent publications include *Wittgenstein and the End of Philosophy* (2006), *Folk Psychological Narratives* (2008) and, with Erik Myin, *Radicalizing Enactivism: Basic Minds without Content* (2013).

Oskari Kuusela is Lecturer in Philosophy at the University of East Anglia, UK. He uses Wittgenstein's later conception of the status of philosophical statements as a response to what he sees as an ascetic tendency in the way that philosophers tend to understand the role of philosophical statements. He argues this leads to an increase in the flexibility of philosophical thought, without a loss in its rigour. His recent publications include *The Struggle against Dogmatism: Wittgenstein and the Concept of Philosophy* (2008) and *The Oxford Handbook of Wittgenstein* (2011, edited with Marie McGinn).

Michael D. Maraun is Professor of Psychology at Simon Fraser University, Canada. The focus of his work is in psychometrics, particularly the latent variable model, the logic of measurement and the special case of construct validation theory. His published work includes articles

concerning the relevance of Wittgenstein's philosophy of psychology and philosophy of mathematics for psychological measurement. His recent publications concern matters ranging from psychological measurement practices to the issue of what it means for a variable to be unobservable.

Danièle Moyal-Sharrock is Senior Lecturer in Philosophy at the University of Hertfordshire, UK. Her work focuses on what she calls 'the third Wittgenstein' (the post-*Investigations* corpus), and in particular *On Certainty*, which she considers to be Wittgenstein's third masterpiece. Her published work includes *The Third Wittgenstein: The Post*-Investigations *Works* (2004), *Understanding Wittgenstein's* On Certainty (2007), *Perspicuous Presentations: Essays on Wittgenstein's Philosophy of Psychology* (2007), and the forthcoming *Hinge Epistemology* (with A. Coliva).

John Preston is Senior Lecturer of Philosophy at the University of Reading, UK. His research interests are in the philosophy of science, of mind and of cognitive science, and in epistemology. His publications include *Wittgenstein and Reason* (2008), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (2002, with M. Bishop), and *The Worst Enemy of Science? Essays in Memory of Paul Feyerabend* (2000, with G. Munévar and D. Lamb). See www.wittgensteinchronology.com for excerpts of his forthcoming book, *Ludwig Wittgenstein – A Chronology of his Life and Work*.

Diane Proudfoot is Associate Professor of Philosophy at the University of Canterbury, NZ. Her current research is in philosophy of language, philosophical logic, history and philosophy of computer science, philosophy of psychology, philosophy of religion, Wittgenstein and Popper. Her published work includes articles on Wittgenstein and the relation between language and thought, Wittgenstein and artificial intelligence and several articles on Alan Turing. She is director, with Jack Copeland, of the Turing Archive for the History of Computing, available at www.alanturing.net.

Timothy P. Racine is Associate Professor of Psychology at Simon Fraser University, Canada. His research interests include the development of social cognition and the role of conceptual analysis and evolutionary explanation in psychology and related disciplines. His publications include *The Shared Mind: Perspectives on Intersubjectivity* (2008, edited with J. Zlatev, C. Sinha and E. Itkonen), and *Moving Ourselves, Moving Others: Motion and Emotion in Intersubjectivity, Consciousness and Language* (2012, with A. Foolen, J. Zlatev and U. Lüdtke).

Severin Schroeder is Reader in Philosophy at the University of Reading, UK. His research interests include the philosophy of Wittgenstein, Aesthetics, the Philosophy of Mind and the work of Schopenhauer. His published work includes *Wittgenstein*: *The Way Out of the Fly Bottle* (2006) and *Wittgenstein lesen*: *Ein Kommentar zu ausgewählten Passagen der 'Philosophischen Untersuchtungen'* (2009). He is also editor of the volumes *Wittgenstein and Contemporary Philosophy of Mind* (2001) and *Philosophy of Literature* (2010).

Joachim Schulte is Visiting Professor of Philosophy at the University of Zürich, Switzerland. His research interests include the philosophy of mind and Wittgenstein. His focus in recent years has mainly been Wittgenstein's middle period. His published work includes *Chor und Gesetz: Wittgenstein im Kontext* (1990), *Wittgenstein: An Introduction* (1992), *Wittgenstein*: *Experience and Expression* (1993) and *Wittgenstein: Leben Werk Wirkung* (2005). In 2009, he translated and edited the fourth edition of *Philosophical Investigations* with P. M. S. Hacker.

Stuart Shanker is Distinguished Research Professor of Philosophy and Psychology and Director of the Milton & Ethel Harris Research Initiative at York University, Canada. His research interests include early child development, developmental disorders and ape language. His published work includes *Apes, Language and the Human Mind* (1998, with S. Savage-Rumbaugh and T. Taylor), *Wittgenstein's Remarks on the Foundations of AI* (1998), *The First Idea* (2004, with S. Greenspan), *Human Development in the Twenty-First Century* (2008, with A. Fogel and B. King), and *Calm, Alert and Learning: Classroom Strategies for Self-Regulation* (2012).

Wes Sharrock is Professor of Sociology at the University of Manchester, UK. He has had a career-long interest in the philosophy of social science, especially the implications of Wittgenstein's philosophy for social science, including the philosophy of mind, involving an opposition to reductionism in all its forms. His publications include *The Philosophy of Social Research* (1997), *Kuhn: Philosopher of Scientific Revolution* (2002, with R. Read), *Garfinkel and Ethnomethodology* (2003, edited with W. Lynch and M. Sage), and *Studies of Work and the Workplace in HCI* (2012, with G. Button).

Kathleen L. Slaney is Associate Professor of Psychology at Simon Fraser University, Canada. Her current research interests include historical and conceptual analysis of methodological approaches within psychological science, philosophy of psychology, and theoretical and applied psychometrics. She was recently awarded the Sigmund Koch Award for Early

Career Contributions to Psychology from the *Society for Theoretical and Philosophical Psychology*. She is editor, with J. Martin and J. Sugarman, of the forthcoming volume *The Wiley Handbook of Theoretical and Philosophical Psychology*.

Michel ter Hark is Dean of Faculty of Arts and Professor of Philosophy of Language at VU University, the Netherlands. His research interests include the history of analytical philosophy, the history of the cognitive sciences, and the philosophy of language. His published work spans topics such as Wittgenstein, Popper, psychological uncertainty and visual uncertainty. His book-length publications include *Beyond the Inner and the Outer: Wittgenstein's Philosophy of Psychology* (1990), and *Popper, Otto Selz and the Rise of Evolutionary Epistemology* (2004).

Michael A. Tissaw is Associate Professor of Psychology at the State University of New York at Potsdam, Unites States. His areas of research focus are theoretical and philosophical psychology, generally, and Wittgenstein and the philosophical analysis of psychological concepts, and early social development, more specifically. Using Wittgensteinian principles, he has published a number of works critiquing the habituation paradigm that is commonly employed infant developmental research. With Rom Harré, he published *Wittgenstein and Psychology: A Practical Guide* (2005).

Leonidas Tsilipakos is a doctoral student in the School of Social Sciences at the University of Manchester, UK. His research interests include Wittgensteinian and ordinary language philosophy, philosophy and methodology of social science, social theory, ethnomethodology and conversation analysis. In his doctoral work, he is exploring the relevance of Wittgenstein and ordinary language philosophy for resolving conceptual problems that arise within the discipline of sociology. His published works have addressed topics such as critical realism and ontological reasoning.

Meredith Williams is Professor of Philosophy at Johns Hopkins University, United States. Her research interests include the philosophy of later Wittgenstein, philosophy of mind and psychology, and the history of experimental psychology. She has written on topics such as the problem of rule following, the importance of learning in Wittgenstein's later philosophy, and the computational theory of mind. Her published works include *Wittgenstein, Mind and Meaning: Toward a Social Conception of Mind* (1999), *Wittgenstein's Philosophical Investigations Critical Essays* (2006), *and Blind Obedience: Structure and Content of Wittgenstein's Later Philosophy* (2010).

Abbreviations of Wittgenstein's Works

AWL *Wittgenstein's Lectures, Cambridge 1932–35*, A. Ambrose (ed.) (Oxford: Blackwell, 1979).

BBB *The Blue and Brown Books: Preliminary Studies for the Philosophical Investigations* (Oxford: Blackwell, 1958).

BEE *Wittgenstein's Nachlass: The Bergen Electronic Edition* (Oxford: Oxford University Press, 2000).

BT *The Big Typescript*, G. G. Luckhardt and M. A. E. Aue (eds and trs) (Oxford: Blackwell, 2005).

CV *Culture and Value*, G. E. von Wright and N. Nyman (eds), P. Winch (tr.) (Oxford: Blackwell, 1980).

LFM *Lectures on the Foundations of Mathematics*, C. Diamond (ed.) (Chicago: University of Chicago Press, 1976).

LPE 'Notes for Lectures on "Private Experience" and "Sense Data"' in J. Klagge and A. Nordmann (eds) *Philosophical Occasions 1912–1951* (Indianapolis: Hackett), 202–88.

LPP *Wittgenstein's Lectures on Philosophical Psychology 1946–47*, P. T. Geach (ed.) (Chicago: University of Chicago Press, 1989).

LSD 'The Language of Sense Data and Private Experience: Notes by Rush Rhees' in J. Klagge and A. Nordmann (eds) *Philosophical Occasions 1912–1951* (Indianapolis: Hackett), 290–367.

LW I *Last Writings on the Philosophy of Psychology*, Vol. 1, G. H. von Wright and H. Nyman (eds), C. G. Luckhardt and M. A. E. Aue (trs) (Oxford: Blackwell, 1982).

LW II *Last Writings on the Philosophy of Psychology*, Vol. 2, G. H. von Wright and H. Nyman (eds), C. G. Luckhardt and M. A. E. Aue (trs) (Oxford: Blackwell, 1992).

MS Manuscripts from the *Nachlass* (References are by MS number followed by page number).

NB *Notebooks, 1914–1916*, G. E. Anscombe and G. E. von Wright (eds), G. E. M. Anscombe (tr.) (Oxford: Blackwell, 1979).

OC *On Certainty*, G. E. M. Anscombe and G. H. von Wright (eds), D. Paul and G. E. M. Anscombe (trs) (Oxford: Blackwell, 1977).

PG *Philosophical Grammar*, R. Rhees (ed.) (Oxford: Blackwell, 1974).

PI *Philosophical Investigations* 2nd edn, G. E. M. Anscombe and R. Rhees (eds) (Oxford: Blackwell Publishers, 1959) [4th edn, P. M. S. Hacker and J. Schulte (eds), G. E. M. Anscombe, P. M. S. Hacker and J. Schulte (trs) (Oxford: Wiley-Blackwell, 2009)].

PI II *Philosophical Investigations Part* II, G. E. M. Anscombe and R. Rhees (eds) (Oxford: Blackwell Publishers, 1959).

PO *Philosophical Occasions 1912–1951*, J. Klagge and A. Nordmann (eds) (Indianapolis: Hackett, 1993).

PPF *Philosophy of Psychology – A Fragment* (Formerly known as Part II of *Philosophical Investigations),* P. M. S. Hacker and J. Schulte (eds), G. E. M. Anscombe, P. M. S. Hacker and J. Schulte (trs) (Oxford: Wiley-Blackwell, 2009).

PR *Philosophical Remarks*, R. Rhees (ed.), R. Hargreaves and R. White (trs) (Oxford: Blackwell, 1975).

RFM *Remarks on the Foundations of Mathematics*, G. H. von Wright, R. Rhees and G. E. E. Anscombe (eds), G. E. M. Anscombe (tr.) (Oxford: Blackwell, 1978).

RPP I *Remarks on the Philosophy of Psychology,* Vol. 1, G. E. M. Anscombe and G. H. von Wright (eds), G. E. M. Anscombe (tr.) (Chicago: University of Chicago Press, 1980).

RPP II *Remarks on the Philosophy of Psychology,* Vol. 2, G. H. von Wright and H. Nyman (eds), C. G. Luckhardt and M. A. E. Aue (trs) (Chicago: University of Chicago Press, 1980).

TLP *Tractatus Logico-Philosophicus*, D. F. Pears and B. F. McGuiness (London: Routledge and Kegan Paul, 1961).

TS Typescript from the *Nachlass* (References are by TS number followed by page number).

VW *The Voices of Wittgenstein: The Vienna Circle*, G. Baker (ed.) J. Connolly and V. Politis (trs) (London: Routledge, 2003).

Z *Zettel,* G. E. M. Anscombe and G. H. von Wright (eds), G. E. M. Anscombe (tr.) (Los Angeles: University of California Press, 1970).

Introduction: Conceptual Analysis and Psychology: An Overview

Timothy P. Racine and Kathleen L. Slaney

I.1 The motivation for and description of the present volume

It has proven surprisingly difficult for psychologists to find unanimous or even unambiguous answers to seemingly simple questions like 'When do infants and children understand intentions or beliefs?' or 'Do primates share intentions with others when they gesture?' or even 'Is this phenomenon best explained by conditioning or high-order cognitive processes?' One reason for this is healthy scientific debate concerning whether a particular gesture or class of gestures truly requires the coordination of intentional behaviour between interlocutors, or whether grounds for belief or other higher-order psychological concepts are satisfied in a particular research theoretical framework. However, another reason, one that we believe to be causing considerable nuisance in contemporary social and behavioural science research, is non-scientific. The root problem is the lack of consideration for the meanings of concepts that are in play in such work and the philosophical positions that are taken, explicitly or otherwise, by the researchers who interpret such psychological terms in particular ways.

We organized the present volume, *A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology*, so that philosophers and scientists could describe and apply a form of conceptual analysis, often associated with the philosophy of Ludwig Wittgenstein, to historical and contemporary psychological empirical research and theory construction. Our expectation was that such a volume would help to clarify the application of particular scientific concepts while providing sufficient background to enable researchers to employ these methods in their particular research areas. A further motivation for the volume was

to contribute to and extend Wittgensteinian scholarship concerning psychology as a discipline. As such, this volume introduces and explicitly discusses the relevance of Wittgenstein and conceptual analysis in general for psychologists. In particular, it provides a description of psychological issues of interest to Wittgenstein and examples of conceptual analytic work – a rough set of methods for use in diagnosing and remediating research problems stemming from unclear or hasty uses of scientific concepts.

Wittgenstein wrote extensively on the philosophical foundations of psychology and the classification of psychological concepts in particular in the last few years of his life (e.g. RPP I, RPP II, PPF, OC). Although his discussions of, for example, William James, Wolfgang Köhler or his allusions to behaviourism or Freud might seem anachronistic to contemporary psychologists, as this volume will show, the ways of thinking that are exposed by his analysis are very much with us today. In fact, the contributors to the present volume, a distinguished group of philosophers and social and behavioural scientists, have all published substantial amounts of work documenting that this is the case. The contributors include Alan Costall, Hans-Johann Glock, Nicole Hausen and Michel ter Hark, Daniel Hutto, Oskari Kuusela, Michael Maraun, Danièle Moyal-Sharrock, John Preston and Severin Schroeder, Diane Proudfoot, Joachim Schulte, Stuart Shanker and Devin Casenhiser, Wes Sharrock and Leonidas Tsilipakos, Michael Tissaw, and Meredith Williams, and the editors, Tim Racine and Kate Slaney. We are particularly privileged that P. M. S. Hacker, one of the world's foremost authorities on Wittgenstein, has provided the prologue to this collection.

It was our intent that this collection of chapters would span a fairly heterogeneous set of topics, thereby demonstrating the relevance that Wittgenstein has for contemporary work in the social and behavioural sciences. The volume begins with Hutto, who makes an impassioned case for the importance of conceptual work in psychology that he proceeds to apply to debates concerning folk psychology, a research area that attempts to account for commonsense concepts of mind such intention, desires and beliefs. Kuusela takes Hutto's plea for conceptual analysis as an invitation to explain, in more general terms, what conceptual analysis is and the issues it raises. The two chapters that follow, by Schulte, and Hausen and ter Hark, respectively, can be thought of as exemplary case studies in method that build upon Kuusela's preparatory work. Schulte's exegetical analysis of two oft-cited fragments concerning the relation between the body and soul demonstrate a familiar theme in Wittgenstein's work, namely the relation between the inner and outer. The centrality of

internal relations is a theme picked up as well in Hausen and ter Hark's Wittgensteinian analysis of aspect seeing in Gestalt psychology.

With exposure to, and demonstration of, Wittgenstein's methods well in hand, Williams compares and contrasts the foundations of mathematics and psychology through an analysis of their respective uses of propositions and norms. Glock continues this expanding of focus to the tricky issue of what we can say of animal minds and argues for a form of 'impure conceptual analysis' that includes empirical and methodological considerations. The theme of mental life continues in Moyal-Sharrock's analysis of John Searle's influential philosophy of mind, with particular attention paid to its causal and representational underbelly. Searle's philosophy is also a topic that is central in Proudfoot's examination of human-computer interaction and the extent to which robots can satisfy the grounds for the attribution of emotional states. The relation between behavioural expression and internal state continues in Sharrock and Tsilipakos' critical analysis of Margaret Archer's theory of internal conversation.

The next five chapters of the volume concern relatively more constrained psychological research areas. Shanker and Casenhiser discuss self-regulation in both a historical and modern context and display current unclarities in contemporary work in this area. Maraun scrutinizes the coherence of the classification of suicide in order to critique confusion between causes and reasons that appears in much of this literature. Preston and Schroeder discuss the extent to which neuroscientific evidence can be marshalled in favour of a representational theory of perception. Racine and Slaney consider social psychological theories that seek to explain why persons find death a fairly terrifying idea. Tissaw reviews popular methods in developmental psychology that are often claimed to show that human infants have complex, yet pre-linguistic, forms of understanding. In the final chapter, Costall conducts a more historical analysis of the same sorts of conclusions and shows why cognitivist forms of explanation remain very influential, despite their reliance on the sort of inner-outer dichotomization that Wittgenstein calls into question.

Because Hacker has detailed, in a very clear and compelling manner, many implications of Wittgenstein's philosophy and methods for psychology in his prologue, in the remainder of this introduction we thought it would be helpful to say a few words for those who know very little about Wittgenstein, or know some but wish to know more. We hope that this will serve as a backbone for the themes developed in the prologue and more generally throughout the book.

I.2 Wittgensteinian conceptual analysis in broad strokes

What is conceptual analysis, and how might it be helpful to social and behavioural scientists? As the volume will suggest, there is not a single or straightforward answer to these questions. However, in our view the most important aspect of conceptual analysis for Wittgenstein is not so much to follow some prescribed set of general methods but rather to development a certain *attitude* to philosophical puzzles, including those that arise in psychology. Not an attitude of suspicion, dogmatism or scepticism, but of curiosity, open-mindedness and scrupulous attention to detail. Although it is common, to speak of Wittgenstein's philosophy as containing a set of methods, which it surely does, we would suggest that these methods are a *consequence* of the Wittgensteinian attitude toward paying careful attention to the concepts that feature in our empirical work and theories.

Many attitudes, however, take time to develop, and there are specific steps one can take if attempting a conceptual analysis of a particular psychological issue. In our view, it might go roughly as follows. Perhaps the first step is to compare uses of the concept(s) in question in order to determine whether researchers are using terms in an ordinary everyday sense, a non-ordinary sense (i.e. using a familiar term in an unfamiliar or perhaps highly restricted manner, but without providing an explicit definition of the alternative use) or a more technical manner (i.e. by means of an explicit technical definition). This is not to lay down rigid rules for the use of the concepts in our languages, but rather to determine what the authors mean by the term(s) employed and how reasonable their claims are in light of their stated goals and/or definitions. For, as Wittgenstein states (PI, §130), 'language-games are rather set up as "objects of comparison" which are meant to throw light on the facts of our language by way not only of similarities, but also of dissimilarities'. (See further discussion by Oskari Kuusela, this volume.) This is not to say that Wittgenstein was much of a pluralist though, for as he remarks later in the *Investigations:*

> [His] aim is: to teach you to pass from a piece of disguised nonsense to something that is patent nonsense. (PI, §464)

The second step will often involve distinguishing atypical or obscure uses of concepts – which will likely have relatively clear theoretical or practical consequences – from idiosyncratic uses on the parts of particular researchers, in which the term in question is used in a casual

and relatively innocuous manner. Obviously, it is far from innocuous to base an entire theory on a foundation of conceptual sand that, in turn, leads to an unclear empirical program of work that ultimately shows up in a feature article in the science section of the *New York Times*. However, it is probably pretty harmless for a researcher studying basic attentional processes to use the concept 'attention' in a restricted manner in the study of, say, eye-tracking behaviour in human children as an end in itself. A final and additional and important step is often to clarify what the researcher seems to mean to say and saying it for him or her. When viewed in this light, Wittgensteinian conceptual analysis seems to be just a commonsense approach to science, which, in our opinion, it is.

I.2.1 But that is not the Wittgenstein I've heard about

Those with some previous familiarity with Wittgenstein, particularly his remarks concerning psychology, might be struggling to reconcile the view they might already have of Wittgenstein with the one just presented. After all, did he not remark that 'The confusion and barrenness of psychology is not to be explained by calling it a "young science"' and that 'in psychology there are experimental methods and conceptual confusion' (PPF, §371)? (See Hutto's chapter in the present volume for further analysis of these remarks and also Hacker in his prologue.) However, it is critical to understand *why* Wittgenstein wrote these things and what he meant by them. Although we will make a few preliminary expansions on these topics presently, we hope that this volume in its entirety will provide the appropriate corrective.

First, Wittgenstein is pointing out, by responding to an implicit claim by Köhler, that the historical origins and research practices of psychology in Wittgenstein's time are not comparable to those of physics, which to our eyes seems hardly a controversial contention. The subject matter of psychology, for the most part, requires the use of a familiar set of everyday concepts (e.g. intention, belief, emotion, thought, helped, cooperated) that have quite complex and criss-crossing fields of use. (In Wittgenstein's terms, they have complicated *grammars*.) For example, the sense in which a psychologist observes another's intentions is conceptually distinct from the way that a physicist observes quarks in a particular accelerator. Now, to be fair, not everyone would agree with this, and some urge that a reduction of psychology to biology and ultimately physics is not only a possible but also a laudable goal. However, as Moyal-Sharrock implies in her contribution to the present volume, there are far-reaching consequences to

whether one agrees that psychological concepts like 'intentions' are ultimately like physical objects. In any everyday sense, they certainly are not. This is not say that they do not in some sense have physical causes, but that is another matter. In either case, in Wittgenstein's view, and indeed in that of the contributors to the present volume, the sorts of discoveries that are made in the social and behavioural sciences are, for the most part, the result of determining the empirical character of familiar phenomena that have pre-existing meanings. Physics, on the other hand, requires a technical vocabulary to represent unexpected objects of inquiry such as 'dark matter' and, more historically, familiar ones like 'electron.'

Second, with respect to the issue of Wittgenstein's anti-theoretical or, perhaps more accurately, non-theoretical attitude, it seems to us that it is a consequence of a form of what might be called 'particularism'. In *The Blue and Brown Books* (BBB, p. 18), he cautions against holding a 'contemptuous attitude towards the particular case.' By contrast, in the social and behavioural sciences the adequacy of an explanation is often understood to be a function of its generality. To see Wittgenstein's point, because it is a topic that shows up in a variety of chapters, we will now consider the example of a 'representation' and briefly discuss what is known as the 'representational theory of mind' (RTM). (For more detail, see chapters in the present volume by Costall, Hutto, Moyal-Sharrock, Preston and Schroeder, Proudfoot, and Racine and Slaney.)

To be exceedingly brief, RTM is essentially a theory or set of theories that posit (mental) intermediaries whose role is to represent to agents, via symbols, the objects or states of affairs that they perceive in the external world. This, then, is a theory in the most general sense; it stipulates that it is the general case that mental representations of the external world *must* occur in order for interaction with the world to occur. The idea of a must – indeed, a metaphysical must – is precisely to hold a 'contemptuous attitude towards the particular case.' By contrast, Wittgenstein would urge that one consider whether the grounds for mental representation are satisfied in the particular case. This brings us full circle to how it is that we say that an agent is, for example, thinking. For Wittgenstein, in third-person cases we typically use the circumstances in which particular behaviours are embedded (Proudfoot, this volume); in first-person cases, it is sincere first-person expressions of thought that are at issue. (For more exposition concerning first- versus-third person asymmetry, see chapters by Glock, Hacker, and Maraun.)

Now, in typical cases, we say someone is thinking because they either express the object of their thoughts or they show that, or even what, they are thinking in how they behave, given an appropriate circumstance. Are we licensed here to speak of representation in some general sense? It is not clear what would be gained. However, there might *particular cases*, where we might be licensed. One example might be someone studying a map and then using it to get around a neighbourhood. Or a person might say they are thinking about their vacation last year when they visited some interesting ruins and are recounting it in some detail. These are mundane cases of mental representation that Wittgenstein would have no trouble with at all. Therefore, if the reader has heard that Wittgenstein is some sort of anti-mentalist, the reader has been mislead. The trouble Wittgenstein would have with mentalism is the *metaphysical must* – the general theoretical claim that such *all* processes of psychological interest can be simplified and unified in this sense. To do so would distort their meaning and alter the very lives that we live. Wittgenstein would also have considerable difficulty with the idea that interesting psychological behaviour is the *necessary causal outcome* of representational processes that seem to be mere redescriptions of the phenomena of interest.

I.3 Conclusion and two cautions

We believe that the set of chapters presented in the collection will provide a spur for some fresh work in psychology and the social and behavioural sciences more generally. We do wish to conclude our brief overview with two cautionary notes for those who are interested in welcoming Wittgensteinian methods of conceptual analysis into their areas of psychological research interest.

First, it is sometimes claimed that Wittgenstein advocates a certain *style* of psychology. For example, Harré and Tissaw (2005) argue that Wittgenstein's philosophy implies a new paradigm that is discursive. If Wittgenstein's goal though were primarily to dissolve philosophical puzzles through adopting a particular attitude and set of methods, why would this directly imply anything about the *general* ontological status of psychology? It certainly would not mean that experimental research is somehow suspect and that discursive work should replace it. (Of course, not all empirical questions are amenable to experimental analysis, and not all phenomena of psychological interest are necessarily measurable either, but this seems a far cry from closing up the laboratory in favour of discursivism.) Hutto (this volume) defends experimental work – for

well-formed questions – and there is ample evidence in Wittgenstein's own work that he had no animus toward experimentation. In our view, at best, these suggestions should be taken to call into question the dichotomization of the inner and the outer that is essentially institutionalized in many, but not all areas, of psychology. However, one should not lose sight of the fact of that the inner *is not* the outer, as is sometimes suggested as well, for example, in some species of distributed cognition; Wittgensteinian conceptual analysis should help us see the internal relations between inner and outer, body and mind, but to help us not collapse the one to the other.

Another way that Wittgenstein is sometimes potentially misconstrued in psychology in our view is by incorporating his ideas into individual psychological theories. This is understandable to some degree because Wittgenstein (Z, §412) asks himself questions like 'Am I doing child psychology?' However, to use Wittgenstein as a *basis* for psychological theorizing is to miss much of the motivation for the present volume. Wittgenstein provides a way of thinking and a set of methods for dissolving philosophical confusions, including those that occur in (child) psychology. And, to return to this quote from *Zettel*, Wittgenstein answers the rhetorical 'Am I doing child psychology?' by claiming he is 'making a connexion between the concept of teaching and the concept of meaning' (Z, §412).

Although the editors of the present volume have learned that it is risky to answer Wittgenstein's rhetorical questions – and even riskier to take them out of context, we feel fairly certain that in making a logical connection between the concepts of teaching and meaning, Wittgenstein is not meaning to engage in or encourage a form of child developmental theorizing. Statements about a child *qua* language learner, or private language arguments, are meant to look at how language functions, much in the same way that Wittgenstein uses primitive language-games; they are not a theoretical statement about how children do/do not learn language or for that matter features of their mentality. Why? This would be to take a logical point and make of it an empirical hypothesis, which is clearly against the spirit of what motivated his work. Although some philosophers are sceptical about the claim that Wittgenstein does not put forth *philosophical* theses, we are quite confident that Wittgenstein does not mean to put forward *psychological* theses.

This is not say that Wittgenstein, like many other important historical figures, might not inspire psychologists to look at things anew and in creative ways that lead to theoretical innovation, but it is quite unclear

what a 'Wittgensteinian psychology' might look like. It is far better, in our view, to take his methods and use them in the way that he intended.

Reference

R. Harré and M. Tissaw (2005) *Wittgenstein and Psychology: A Practical Guide* (Aldershot, UK: Ashgate).

Prologue: Wittgenstein's Philosophy of Psychology as a Critical Instrument for the Psychological Sciences

Peter M. S. Hacker

1 Wittgenstein's concern with the philosophy of psychology

Wittgenstein's interest in the philosophy of psychology began after his return to philosophy in 1929. As he became aware of the grave mistakes he had made in the *Tractatus*, he came to realize that one of the roots of the illusions that had beset him lay in the particular form of anti-psychologism that he had taken over from Frege.[1] In the *Tractatus,* he had written:

> Psychology is no more closely related to philosophy than any other natural science. Does not my study of sign-language correspond to the study of thought processes, which philosophers used to consider essential to the philosophy of logic? Only in most cases they got entangled in unessential psychological investigations, and with my method too there is an analogous risk. (TLP, 4.1121)

The result was that he failed to give due attention to meaning (*meinen*), intending, thinking, understanding, interpreting and knowing. These had all been brushed under the carpet in the *Tractatus* on the misguided Fregean grounds that they are the concern of psychology, not of philosophy. This meant that Wittgenstein failed to examine his own presuppositions concerning what it is to mean something by a word one uses or a sentence one utters, or what it is to know what a word means or to understand what has been said, or what it is to communicate with another or to interpret the words of another. For to be sure, one cannot

discuss the nature of representation, in general, or sentences with a sense, in particular, without some presuppositions concerning what it is to understand an utterance or to know the meanings of expressions, or to mean something by the use of an expression. These are, one may think (and Wittgenstein thought), mental processes of some kind (such as understanding), or perhaps mental states (such as knowing), or maybe mental acts (such as meaning something by a word). And now, as he realized only much later,

> The first step is the one that altogether escapes notice. We talk of processes and states, and leave their nature undecided. Sometime perhaps we'll know more about them – we think. But that's just what commits us to a particular way of looking at the matter. (The decisive movement in the conjuring trick has been made, and it was the very one that seemed quite innocent). (PI, §308)

As the *Tractatus* began to unravel, Wittgenstein came to realize that all these psychological attributes are *internally related* to linguistic representation. For the meaning of a word is what one knows when one knows what a word means, and the meaning of an utterance is what one understands when one understands what someone has said. Although not every utterance could possibly stand in need of an interpretation, when an utterance can be understood in more than one way, an interpretation is called for. When one utters a sentence containing an indexical, one means by it the item one uses it to refer to, and when one utters a sentence to describe a state of affairs, one means by the sentence the state of affairs one thereby describes. It is precisely because these are internal relations that the *concepts* require philosophical scrutiny. For internal relations are *constitutive* of *both* relata. The investigation of these concepts provided one of Wittgenstein's two main routes into philosophy of psychology. This one occupied Wittgenstein from 1930 until the completion of the *Investigations*. The other route was via the problems of solipsism.[2] His reflections on solipsism came to maturity in the dictations of *The Blue Book*, and it was these that paved the way for the great private language arguments in the *Investigations*. These, in turn, provided the seedbed for his later reflections of the philosophy of psychology.

For the most part, Wittgenstein's general methodological principles and over-all conception of philosophy were fully formed by the mid-1930s. These principles of 'theoretical philosophy' (to use Kant's terminology)[3] are the following:

- Philosophy is not a cognitive discipline. Philosophy is a contribution to human understanding, not to human knowledge (in the sense in which the natural sciences contribute to human knowledge).
- Philosophy is purely descriptive. What it describes or states are rules for the use of words. These are familiar to anyone who has mastered their use. What philosophy does is to *remind* us of our own usage, to *select* the relevant array of sense-determining rules, and to *order* them in such a manner as to shed light upon the conceptual problem or problems at hand.
- Philosophy is an *a priori* investigation for purposes of conceptual clarification. It aims to resolve or dissolve conceptual confusions and entanglements, and to clarify conceptual relationships.
- There are no theses in philosophy. What look like theses are no more than grammatical propositions in the misleading guise of descriptions of *de re* necessities. Grammatical propositions are norms of representation, rules for the use of their constituent expressions.
- There are no theories in philosophy on the model of hypothetico-deductive theories in the sciences. Grammatical propositions are part of the web of words, but not part of any theory. Although they can be said to be true (just as it is true that the chess-king moves one square at a time), they are not *hypotheses* to be confirmed or infirmed by experience or experiment.
- There are no explanations in philosophy on the model of causal explanations in the natural sciences. The only explanations (clarifications) are grammatical.
- Descriptions of grammar and explanations of conceptual similarities and differences *leave grammar as it is*. It is not the business of philosophy to change or to recommend changes to existing grammar, but only to clarify it as it is. It is the task of the sciences to introduce new terminology for their theoretical purposes, but philosophy has no theoretical purposes.

There is much to be said both for, and about, each of these claims, and about their status. Some of the questions are addressed in this volume of original essays edited by Timothy Racine and Kathleen Slaney in the methodological papers by Daniel Hutto: 'Psychology's Inescapable Need for Conceptual Clarification', and by Oskari Kuusela: 'Wittgenstein's Method of Conceptual Investigation and Concept Formation in

Psychology'. Both are concerned with applying Wittgensteinian methods of conceptual clarification to psychology. Further elaboration is given by Hanjo Glock in his paper 'Animal Minds: Philosophical and Scientific Aspects', in which he discusses Wittgenstein's methodology with special reference to the description of animals and animal behaviour. These general principles, all of which characterize the *Investigations*, remain firmly in place throughout Wittgenstein's writings on philosophy of psychology. They are, of course, perfectly consistent with Wittgenstein's drawing our attention to very general facts of nature that provide the background for our language-games. This is a leitmotif that becomes more evident in the 1940s but in no way undermines his methodological principles. Such general facts of nature are drawn to our attention in order to make it clear that our concepts are not right or wrong and are not reflections of the nature of things, even though they are conditioned by our nature and the nature of the world in which we live. Were these different in imaginable ways, our conceptual scheme would differ.

The central preoccupation of the *Investigations* is with language and linguistic representation. The first part of the book (§§1–189) is Janus-faced. Here Wittgenstein confronts his old views, digs down to their roots and examines the unquestioned presuppositions that he had shared with the great tradition of European philosophy. He criticizes his earlier conceptions and investigates the same problems *de novo*. His new solutions, resolutions and dissolutions of these problems overthrow centuries of philosophical reflection. Originally, Wittgenstein had meant to continue with what is now Part 1 of the *Remarks on the Foundations of Mathematics*, that is with his new account of the nature of logical and mathematical necessity. However, in the early 1940s, he changed his mind. He replaced his reflections on the nature of necessity with a continuation of his remarks on following rules and practices, and then embarked on the private language arguments.[4] These are no less relevant to the philosophy of language than they are to the philosophy of psychology. Thereafter, much of the book is concerned with themes in philosophy of psychology, such as thought, imagination, rationality, memory, the self, consciousness, intention and meaning something by an expression. However, these are all linked more or less firmly with linguistic meaning and representation. It is evident that problems in the philosophy of psychology increasingly captured his imagination and stimulated his thought.

2 Wittgenstein's revolution in the philosophy of psychology in the *Investigations*

Although Wittgenstein's concern with psychological concepts in the *Investigations* was thus constrained by the over-all project of the book, it is evident that he brought to light some of the most fundamental logico-grammatical features of psychological expressions. These insights provided the framework for his subsequent unconstrained reflections on psychological concepts in the years 1946–1950.

The fundamental principles (grammatical insights) of Wittgenstein's philosophy of psychology that are laid forth in the *Investigations* can be summarized as follows:

- The subject of psychological attributes is not the ego, the mind or the brain that a sentient being has, but the creature as a whole.[5]
- The conception of experience as privately and inalienably 'owned' by the subject of experience is misconceived. Different people can have the same experience (neither qualitatively the same, nor numerically the same, but just the same), and often do.
- The conception of experience as epistemically private, i.e. known only to the subject is doubly misconceived. We very often do know how others are feeling and what they are thinking, and we cannot be said to know or to be ignorant of how we are feeling or what we think.
- There is a paramount difference between the first- and third-person use of a range of important psychological verbs. Their first-person present tense is groundlessly applied (*not* on the basis of introspection). Their third-person present tense is applied on the grounds of defeasible behavioural criteria. These are two sides of the same conceptual coin. Unless one has a grip on both, one has not grasped the meaning of the psychological predicate.
- A criterion for the application of a psychological predicate is not inductive evidence, but logically good evidence. Ascription of psychological predicates to others is not typically inferred from the satisfaction of the criteria for their application, but it is warranted by those criteria. The intelligibility of groundless application of such predicates to oneself depends upon their logical connection with behaviour and circumstance that warrant their third-person ascription. Without such logical connections, the intelligibility and immediacy of the first-person use would require a private ostensive definition.
- There is no such thing as a private ostensive definition. Hence too, there is no such thing as the memory of an experience fulfilling the

role of a defining sample. So the meaning of a psychological predicate cannot be explained either for oneself or for others by reference to a private ostensive definition or by reference to a mental sample, (there is no such thing as such an explanation). For neither experiences nor the memory of experiences can (logically) fulfil the role of samples for the application of a word.

- There is a fundamental epistemological difference between the first- and third-person uses. The first-person use of such psychological verbs does not admit of the standard uses of epistemic operators. 'I know I am in pain' is disanalogous to 'I know he is in pain' and to 'I know I was in pain' – all it amounts to is 'I really am in pain'. 'I know what I think' is unlike 'I know what you think' – all it amounts to is either 'I have an opinion on the matter' or 'I have made up my mind on the matter'. For in all such cases, there is no such thing as *not knowing*, *being ignorant*, *having or lacking grounds*, *doubting*, *wondering*, *believing or thinking*, *guessing* or *being mistaken*. By the same token, there is no such thing as *knowing*, *being right*, *being sure or certain*, and *finding out*, *learning* or *detecting*.
- Although there is such a thing as introspection, it is not a kind of inner sense (or 'apperception' as Leibniz denominated it). It is either a form of self-reflection characteristic of introspective personalities like Proust, or a matter of registering how things are with one (as when one keeps a diary of one's pains for the doctor).
- The limits of thought are the limits of the behavioural expression of thought. A creature can intelligibly be said (truly *or falsely*) to think that things are so just to the extent that its behavioural repertoire includes such forms of action and response that would warrant the ascription to it of *thinking things to be so*. The limits of what a non-language-using animal can be said to think are the limits of its cognitive expressive behaviour.

These grammatical principles constitute a far-reaching overview of the grammar of psychology. They are the result of a careful selection, systematic arrangement of, and methodical reflection on, familiar rules for the use of psychological expressions. In each case, the consequence of denying or repudiating these grammatical connections is shown to lead to incoherence. Collectively, these principles revolutionize our philosophical understanding of that segment of our conceptual scheme that is concerned with the psychological.

Given these logico-grammatical insights, the whole Cartesian and post-Cartesian conception of the mental is over-turned. With it go its

degenerate forms, such as idealism (Cartesianism without its materialist spouse), materialism (Cartesianism without its mental spouse) and its various offshoots, such as identity theory (type as well as token-identity theories), functionalism, as well as behaviourism and logical behaviourism, and the brain/body dualism that characterizes contemporary cognitive science and cognitive neuroscience. In effect, the map of possibilities envisaged by philosophers, psychologists, and neuroscientists over the last four centuries has to be scrapped and replaced by a quite different one. Wittgenstein's further efforts at mapping the logical geography of the mind were developed in his post-*Investigations* writings.

3 Wittgenstein's remarks on the philosophy of psychology: methodological guidelines

In 1946, with a clear philosophical methodology and with a firm grasp of the fundamental logico-grammatical characteristics of psychological concepts, Wittgenstein was able to concentrate his efforts on the philosophy of psychology. His purpose was to give a rough sketch of the landscape of the mind. The point of all his classifications and comparisons, he said, was that they can answer a whole battery of philosophical problems in this domain. For they demonstrate a *method* of getting clear about conceptual difficulties in the domain of the mental (MS 134, p. 156). 'The philosopher wants to master the geography of concepts; to see every locality in its proximate and also in its most distant surroundings' (MS 137, p. 63a). Indeed he pushed the cartographical metaphor further:

> In order to know your way about an environment, you don't merely need to be acquainted with the right path from one district to another; you need also to know where you'd get to if you took the wrong turning. This shows how similar our considerations are to travelling in a landscape with a view to constructing a map. And it is not impossible that such a map will sometime get constructed for the regions we are moving in. (MS 131, p. 121 = RPP I, §303)

'The difficulty', he wrote a little later, 'is to know one's way among concepts of "psychological phenomena". To move about them without repeatedly running up against an obstacle. That is to say: one has got to *master* the kinships and differences of concepts' (MS 135, p. 73 = RPP I, §1054). For to have mastered the use of an expression, to possess the concept expressed by a term, is not to have mastered its *comparative*

use. But that is precisely what is needed when confronted by conceptual confusions.

In the course of his reflections, Wittgenstein elaborated a number of crucially important guidelines for investigations of psychological concepts. I shall briefly state them, (space does not permit their defence).

- The general categorial expressions philosophers, psychologists and cognitive neuroscientists are prone to deploy are *mental state*, *mental process*, *mental event*, *mental act*, *mental activity*, and *experience*. These categorials seem to be the hardest of the hard, but in fact they are elastic and relatively indeterminate. Moreover, although they appear to be species of a genus, the other species of which is *physical state, physical process, physical event,* and so forth, that is altogether mistaken. A mental state is not 'just like a physical state only mental'.
- Classifying psychological attributes under one of these general categorials is likely to be exceedingly misleading. First, numerous psychological expressions have multiple centres of variation, so that in some occurrences they may signify an occurrent mental state, in others a dispositional mental state, and in yet others a mental activity. Secondly, some psychological concepts, such as *belief*, cannot be subsumed under any useful categorial – believing something is neither a mental act, activity or experience, nor a mental state, process or event.
- The moral of the tale is that if one asks, 'Is this a mental state [or process, activity, act or experience]?', one sees that neither an affirmative nor a negative answer helps. There are too many things that could all be called by this categorial name – and the classification no longer helps. Instead, *one must distinguish the concepts from one another individually* (MS 167, p. 6). This rules out mechanical pigeon-holing, which is the bane of philosophy.
- Rather than succumbing to the temptation to pigeon-hole a problematic psychological concept, examine the *language-games in which it is at home*, the *behaviour with which it meshes*, and the *occasions in which it is appropriate to invoke it*. So too, to combat misleading and baffling surface-grammatical similarities between concepts, for example, between seeing, seeming to see, imagining and dreaming, compare and contrast the language-games in which they occur.
- Always ask how a given psychological concept might be taught, for the way in which such a concept might be taught sheds light on its role. If, for example, one is baffled by the 'indistinguishability' of

perceiving that something is so and dreaming that it is so, start by reflecting on how one might teach a child the use of 'I dreamt'. If one is trapped in the misleading grammar of 'intend' (and perhaps wonders how one knows what one intends, or thinks, like Benjamin Libet, that intentions are felt), it is helpful to consider how, and in what language-games, one might teach a child the use of the phrase 'I'm going to' (namely: as heralding an action, not as the expression of a feeling).

- In order to discern conceptual differences and kinships, one must examine linguistic usage with care. What is needed is not decompositional analysis, but what Peter Strawson was later to call 'connective analysis', that is, the delineation of tracing out the manifold connections of a given problematic expression with adjacent expressions in the web of words. Most psychological concepts do not readily lend themselves to analytic definition, and even when they do, the definition by itself does little to resolve the conceptual problems that irk us – for it is the topography of the landscape that we need to see.
- Many psychological concepts (e.g. consciousness, belief, thinking) have a number of *different* 'centres of variation'. Rather than squeezing such a concept into a distorting pattern, one must describe the concept as it is, with all its unruliness and irregularity. Beware of the temptation to 'tidy up' existing usage. For what one usually does is not to clarify a concept but to falsify it. Special philosophical meanings for familiar expressions (such as 'consciousness', 'belief', or 'knowledge') are derived from the ordinary meaning by a series of misunderstandings and misrepresentations. They result not in a special philosophical sense of 'consciousness', 'belief' or 'knowledge', but only in a special philosophical confusion.
- Always pay close attention to the first-/third-person asymmetries between psychological verbs. The presence of such asymmetries is always highly instructive from a conceptual point of view (and their absence is no less significant).

These principles provide one with substantial methodological guidelines by means of which one can find one's way through the jungle of our psychological concepts. Moreover, having them sharply in mind when reading the two volumes of Wittgenstein's *Remarks on the Philosophy of Psychology* is of no small assistance in understanding the way in which Wittgenstein hacks his way through the undergrowth.

Wittgenstein's reflections on philosophy of psychology that were written after he had completed the *Investigations* run to 1,900 pages

of MSS. There is a great deal there that needs digesting, there is much that is puzzling and there is nothing that is not of interest. Some of the papers in the following volume are concerned with clarifying what Wittgenstein meant by some of the things that he wrote, and how his reflections compare with those of some contemporary philosophers. Joachim Schulte sheds light on the tantalizing remarks: 'My attitude towards him is an attitude towards a soul. I am not of the opinion that he has a soul' (PPF, §22) and 'The human body is the best picture of the human soul' (PPF, §25). Diane Proudfoot shows how much mileage there is in Wittgenstein's remark that a smiling mouth smiles only in a human face (PI, §583) and in subsequent observations on the expressive nature of smiles. She then goes further and instructively applies Wittgenstein's insights to the field of social robotics. Daniele Moyal-Sharrock compares Wittgenstein with John Searle on conceptual elucidation and the pursuit of realism without empiricism. She nicely brings Wittgenstein's reflections in *On Certainty* to bear on fundamental questions in philosophy of psychology and cognitive neuroscience.

4 Bringing Wittgenstein's insights to bear upon the sciences of the mind

Although Wittgenstein has sometimes been misinterpreted as a philosophical quietist,[6] nothing could be further from the truth. For it is Wittgenstein who, for the first time in the history of our subject, has explained why philosophy has *a license* to interfere in the sciences. For scientists are no less liable to conceptual confusions than anyone else, and scientific theorizing is as liable to conceptual entanglement as any other intellectual endeavour. To say that philosophy has a license to interfere in the sciences is not to say that philosophy is *in competition* with the sciences – that philosophy discovers truths about the world as the natural sciences do, and that its claims compete with scientific ones. Nor is it to say that philosophy discovers different truths about the world than do the natural sciences, for example, that while physicists discover empirical truths about this world, *meta-physicists* discover *meta-physical truths* – truths that hold in all possible worlds. There are no 'metaphysical truths' and there is no branch of human knowledge that can be deemed to be *metaphysical knowledge*. What looks like the scaffolding of all possible worlds is no more than the conceptual scaffolding *from which* we describe the actual world. Putative metaphysical propositions are grammatical propositions in misleading guise, or they are confused recommendations for a different grammar, or, more

likely, they are mere nonsense, that is sequences of words that make no sense. The license that philosophy possesses to intervene in scientific debates is a critical one, but the licit criticism is not empirical. It is purely conceptual.

Philosophy is Janus-faced. On the one hand, its task is the description of segments of our conceptual scheme. It aims to give us an overview of the conceptual landscape. On the other hand, its role is to confront conceptual problems and identify conceptual confusions. This twofold *raison d'être* is exemplified both in the direct confrontation with problems in philosophy, and with the identification and rectification of conceptual entanglements in the natural, social and psychological sciences. Philosophy has no title to tell scientists what experiments they should conduct, what theories they should develop, or whether the theories they have developed are true. Its sole title is to examine the questions natural scientists ask in order to see *whether the questions make sense*, to investigate the scientist's *interpretation of the results of his experiments* and to scrutinize the consequent *explanatory theories*, which are meant to answer the questions, to see *whether these theories make sense*. Philosophy is not a conceptual policeman, whose duty it is to stop people from doing things they want to do. It is a conceptual tribunal, whose role it is to tell people whether their questions, assumptions and conclusions make sense (not whether they are true or false), and, if they make no sense, to explain *why* they make no sense. These tasks are of paramount importance in the human sciences, in particular psychology and cognitive neuroscience, but also pedagogy, sociology and economics. For a variety of deep reasons, all these sciences are *especially* liable to conceptual confusion.

Wittgenstein did not engage extensively in the examination of the conceptual flaws in the mental sciences. In his times, with the exception of Freudian psychology, they had nothing like the high profile they currently enjoy. Wittgenstein's interests in the empirical psychology of his day was, for the most part, limited to Freudian psychoanalysis (which dominated the era) and Köhler's *Gestalt* psychology, which was advanced against the prevailing behaviourism of the inter-war years. Wittgenstein knew something of behaviourism, perhaps through Russell's *Analysis of Mind*. He had read James's *Principles of Psychology* (1890) with great care, and to that extent was well informed about work in experimental psychology at the turn of the century, including physical and physiological psychology. It is noteworthy, however, that he used James's book not as a handbook of psychology, but as a 'goldmine' of conceptual confusions in the subject.[7]

With regard to behaviourism, Wittgenstein had clarified his ideas in the 1930s, and had no reason to return to the subject later. He thought that it was a misconception not unlike finitism and formalism in the philosophy of mathematics: 'Finitism and behaviourism are quite similar trends. Both say: but surely, all we have here is ... Both deny the existence of something, both with a view to escaping from a confusion' (RFM, 142). In one of his lectures he remarked, 'We might say that formalism in mathematics is behaviourism in mathematics' (LSD, 111). For formalism and behaviourism invite similar responses: 'Surely numbers are not numerals, but something else' and 'Surely pain is not just behaviour, but something else.' In both cases, the response is misleading. It leads the mathematician to suppose that mathematical propositions describe relations between mathematical objects and that they are true in virtue of their correspondence to a mathematical reality. It leads psychologists to suppose that their subject matter consists of inner experiences directly known only to the subject himself (by introspection), that is that 'psychology treats of processes in the mental sphere, as physics does in the physical' (PI, §571).

Nevertheless, behaviourism was right about *some* matters. *Logical* behaviourism (Carnap and Feigl in the 1930s) was right to insist that there is an internal relation between mental attributes and behaviour. For the criteria for ascribing mental attributes to others consist in their behaviour in the circumstances of life. Where it was wrong was to suppose that the mental is *reducible* to behaviour and dispositions to behave. Ontological behaviourism (Watson and Skinner) was right to emphasize that language learning is based on training, and that it presupposes common behavioural reactions and responses. It was right to conceive of language learning as learning new forms of behaviour – learning how to do things with words. It was correct to conceive of understanding in terms of abilities and dispositions rather than as a hidden mental state or process. But the behaviourists were sorely mistaken to suppose that the mental is a fiction. One can think and feel without showing it, and one can exhibit thoughts and feelings without having them. Avowals of experience are, indeed, a form of behaviour, but what they avow is not behaviour.

With regard to Köhler, Wittgenstein confronted his conception of psychology as a young science, both in the *Investigations* and in *Philosophy of Psychology – a Fragment*.[8] Köhler held that the difficulties that contemporary psychology (unlike the physical sciences) encountered centred on the ineliminability of qualitative judgements of experience. While Köhler repudiated introspective psychology (e.g. of Wundt,

Titchener, and James), he held that the behaviourists had gone too far in supposing that the direct experience of the subject was eliminable from scientific psychology (either on the spurious grounds that there is no such thing [Watson] or on the grounds that such reports are not inter-subjectively verifiable, and hence unsuitable for science [Tolman, Hull]). But, Köhler remonstrated, the direct experience of the subjects remains the raw material, the 'objective experience' of the observational psychologist.[9] However useful pneumographic, plethysmographic and galvanographic methods of measurement may be, they are no substitute for the observational identification of fear, anger or anxiety. The identification of qualitative types of behaviour is indispensable in current psychology, because psychology is still a young science. Physics, a mature science, has eliminated the need for direct qualitative observations of, for example, felt thermal qualities or perceived light intensities in favour of indirect quantitative measurement of temperature and photometric methods of measuring light intensity. Psychology is not yet in the position to do so, for it is, according to Köhler, still in its infancy. In the fullness of time, however, psychology, too, will be able to replace qualitative judgements about mental attributes by quantitative measurements of neurological correlates.

Köhler's conception sounds very congenial to the modern neuroscientific ear – after all, such eminent neuroscientists as Christof Koch have declared that the 'characterizing the NCC [neural correlates of consciousness] is one of the ultimate scientific challenges of our times', a judgement with which his late colleague Francis Crick concurred.[10] Nevertheless, the supposition that the difficulties in psychology, either when Köhler was writing or today, stem from the fact that it is an infant science, let alone from the fact that neural correlates of consciousness have yet to be discovered, is precisely what Wittgenstein challenged more than half a century ago.

The difficulties that beset psychology do not stem from ignorance of neural correlates of consciousness. After all, it is not as if the regularities and functional dependencies that psychology has discovered must be regarded as provisional pending further discoveries concerning conjectured neural correlates. To be sure, the measurement of temperature made it possible for thermodynamics to abandon talk of perceived thermal qualities altogether. But no one (other than misguided philosophers) would suggest that if we were to discover neural correlates of some psychological phenomena, psychology could immediately abandon all talk of perception and sensation, of feelings and emotions, of fearing that *p* and longing for *q*, of deciding

to *V* and intending to *X*, and jettison all existing psychological knowledge associated with these phenomena. After all, these phenomena are the very subject matter of psychology. With their elimination, one would eliminate the very explananda of the science of psychology. The difficulties of psychology are not comparable to those of physics in its infancy. Rather, they are *conceptual* difficulties. That is why Wittgenstein wrote:

> The confusion and barrenness of psychology is not to be explained by its being a 'young science'; its state is not comparable with that of physics, for instance, in its beginnings. (Rather with certain branches of mathematics. Set theory.) For in psychology, there are experimental problems *and conceptual confusion*. (As in the other case, conceptual confusions and methods of proof)
>
> The existence of the experimental method makes us think that we have the means of getting rid of the problems which trouble us; but problem and method pass one another by. (PPF, §371, emphasis original)

Wittgenstein elaborated what he had in mind:

> 'Thinking is an enigmatic process, and we are a long way off from complete understanding of it.' And now one starts experimenting. Evidently without realizing *what* it is that makes thinking enigmatic to us.
>
> The experimental method does *something*; its failure to solve the problem is blamed on its still being in its beginnings. It is as if one were to try to determine what matter and spirit are by chemical experiments. (RPP I, §1093, emphasis original)

In short, it is characteristic of psychology (and today one might add: and of cognitive neuroscience) to rush into experimentation before being in the slightest bit clear exactly what is meant by the expression under which the phenomena to be investigated are subsumed. So leading figures in these subjects tell their audiences that 'Here (pointing at an fMRI scan) one can, *for the first time*, actually see thinking going on' – while being quite oblivious to the fact that whatever may go on in the brain while one is thinking is not the thinking, and that whenever one looks at the living equivalent of *Le Penseur*, one does indeed see thinking going on. For that is the sense we have given to the words 'seeing thinking going on'. Other scientists tell their subjects to 'rotate

mental images in their mind' and try to work out the laws of the movement of mental images in mental space, while being oblivious to the fact that while one can imagine something rotating, one cannot rotate something imaginary – that is there is no such a thing.[11]

Wittgenstein was interested in another important subject discussed by Gestalt psychologists in general and by Wolfgang Köhler in particular, namely aspect perception. This was the object of much of Wittgenstein's writing on the philosophy of psychology after 1946, although the topic had already caught his attention in the *Tractatus* remarks on the Necker cube. I shall say no more about this splendid array of problems, since the subject is discussed in detail in the essay 'Aspect Seeing in Wittgenstein and in Psychology' by Nicole Hausen and Michel ter Hark. They show vividly how Wittgenstein's reflections bring to light conceptual deficiencies in empirical psychology.

The realisation of the importance of applying Wittgenstein's analytic methods to experimental psychology, cognitive science, and cognitive neuroscience dawned late. It was pioneered by Anthony Kenny in his *Action, Emotion and the Will* (1963), but relatively few philosophers took up the challenge or realized the importance of meeting it. But today, in an era obsessed with cognitive neuroscience, fascinated by fMRI scans and what they are alleged to show us about ourselves, and keenly attentive to computational psychology (masquerading as cognitive psychology), the need for methodical conceptual investigation should be evident to anyone concerned with conceptual questions in the sciences of the mind. In an era in which the educated public is bombarded with putative discoveries that are frequently of questionable intelligibility,[12] the importance of careful conceptual scrutiny of such alleged discoveries and of their accompanying theories is surely obvious. Such philosophical investigations into the sciences of the mind and of human behaviour will reveal numerous conceptual unclarities in the formulation of questions, and multiple incoherences in the description of the results of experimentation and in the inferences drawn from them. This, if understood by experimental scientists, will enable psychologists and cognitive neuroscientists to formulate clearer and more fruitful questions, to devise more appropriate experiments, and to construct more adequate explanatory theories.

A substantial number of the papers in this collection, like Diane Proudfoot's mentioned above, are concerned with applying Wittgenstein's ideas to issues in current empirical psychology, cognitive science, artificial intelligence studies and cognitive neuroscience. This is a very welcome shift of attention from interpretation to application. For it is high time

that Wittgenstein's legacy was put to good use in critical conceptual scrutiny of the work being done in the mental sciences. For it is here that philosophy can make an important contribution to science, and clear away conceptual confusions in which scientists are enmeshed and with which they mesmerize and delude the educated lay public. Philosophical investigations will sometimes have direct bearing on practical matters, such as the treatment of autistic children. This is made evident in the paper by Stuart Shanker and Devin Casenhiser 'Reducing the Effort in Effortful Control', which is a prolegomenon to new research on autism and its treatment that is of the highest importance. Whatever is wrong with autistic children, it certainly is not lack of a theory of mind, as Alan Costall shows nicely in his essay 'The Unconscious Theory in Modern Cognitivism'. However perception is to be empirically explained, it is not going to be by means of a representative theory of perception, as John Preston and Severin Schroeder elegantly demonstrate in their chapter 'The Neuroscientific Case for a Representative Theory of Perception'. Their comprehensive criticisms of such eminent neuroscientists as John Smythies, Vilayanur Ramachandran and Chris Frith are definitive. The problematic nature of infant habituation research and the perils of unthinkingly extending the concept of a concept to non-language-users is carefully argued in Michael Tissaw's chapter 'A Conceptual Investigation of Inferences drawn from Infant Habituation Research'. Tissaw offers powerful criticisms of Renée Baillargeon's developmental research that ascribes sophisticated conceptual skills to the brains of neonates and of small children. Wes Sharrock and Leonidas Tsilipakos illuminatingly apply Wittgenstein's reflections on 'the inner' and 'the outer' to Margaret Archer's sociological research on agency and structure. Their essay 'A Return to "The Inner" in Social Theory: Archers' "Internal Conversation"' carefully examines the use and misuse of the notions of talking to others, talking to oneself, talking to oneself in the imagination, and thinking. Michael Maraun turns to psychological and sociological investigations of suicide in order to show the direct relevance of Wittgenstein's observations on the nature of voluntary and intentional behaviour. Michael Maraun's paper 'The Concepts of Suicidology' shows how unclarity concerning the differences between reasons, motives and causes, between objects of, and causes of, emotions, between what is voluntary and what is intentional vitiate the empirical study of suicide. The editors of this volume, Tim Racine and Kathleen Slaney, subject the much-invoked concept of 'psychological meaning' to critical scrutiny in their essay 'Terror Management, Meaning Maintenance, and the Concept of Psychological Meaning'.

All these papers illuminate the ways in which conceptual investigations inspired by Wittgenstein's methods can shed light on scientific theories in the domain of the sciences of the mind. The upshot is anything but trivial. In some cases, scientific theories are shown to make no sense. In other cases, scientific theories are shown not to have the sense that they are thought to have. And in yet others, scientific theories are shown not to have the implications that scientists think they have. To bring clarity to the sciences of the mind and to eradicate conceptual confusions in these sciences is a worthy vocation for philosophy. This volume of essays undertakes this task. It is to be hoped that it will set the model for many further such endeavours.

Notes

1. Frege's conception is well exhibited in a passage from his 'Logic' of 1897/1979, p. 145: 'grasping (*erfassen*; understanding)…is a mental process! And this process is perhaps the most mysterious of all. But just because it is mental in character, we do not need to concern ourselves with it in logic. It is enough that we can grasp thoughts and recognize them to be true; how this takes place is an independent question'.
2. Wittgenstein had embraced a variant of solipsism in the *Tractatus* 5.6–5.641. I have characterized it as 'transcendental solipsism', since it is evidently meant to be consistent with empirical realism (see Hacker, 1986, ch. 4).
3. Theoretical philosophy stands in contrast to *practical philosophy*, such as moral, legal and political philosophy. Wittgenstein's remarks on the nature and limits of philosophy are best suited to those branches of philosophy with which he was concerned, all of which belong to theoretical philosophy in Kant's sense of the term.
4. How it was possible to graft two such different trees onto the same stock is discussed in detail in Baker and Hacker (1985) and in an essay entitled 'Two Fruits upon one Tree' in the extensively revised edition of that work (Baker and Hacker, 2009). Meredith Williams's essay below: 'Parallels in the Foundations of Mathematics and Psychology', further explores this theme.
5. With the exception of verbs of sensation, such as 'to hurt', 'to itch', 'to tickle', which can be ascribed to the body and its parts – as when we say that our foot hurts, itches or tickles. But it is not the foot that has a pain or has an itch or a tickle, it is the person.
6. The source of the misconception is the misunderstanding of his remark that philosophy 'leaves everything as it is' (PI, §124). However, 'everything' here is *everything in grammar*, that is, it is not the task of philosophy to advance a novel and logically improved language (as Frege, Russell and Carnap had supposed or advocated). That, of course, does not mean that philosophy has no substantive consequences. It eradicates nonsense, in both the natural and the human sciences. Philosophy is not the Queen of the Sciences; nor is it a conceptual scullery maid. It is a Tribunal of Sense.

7. In MSS 165, p. 150, Wittgenstein wrote: 'How important work in philosophy is shown by James's *Psychology*. Psychology, he says is a science, but he barely discusses any scientific questions. His movements are only so many attempts to free himself from the spider's web of metaphysics in which he is trapped. *He cannot yet walk or fly at all, he only wiggles*. Not that this is not interesting. It's just that it isn't a scientific activity.' (My translation; italicized sentence written in English.)
8. For detailed discussion, see Hacker (1996), 'Methodology in Philosophical Psychology', section 1.
9. See Köhler's (1929), pp. 28–30.
10. See Koch (2004), p. xvi, preface by Francis Crick.
11. For more extensive examination of Wittgenstein's objections to Köhler's methodological observations, see Hacker (1996), 'Methodology in philosophical psychology'.
12. Such as the 'discovery' that the window of opportunity for the exercise of one's free-will is precisely 150 milliseconds prior to moving; or the 'discovery' that we recognize things by conjuring up a mental image and matching it to what we see (in the course of which we rotate it at constant velocity until it matches); or the 'discovery' that sufferers from blindsight enjoy visual sensations, but cannot convert them into visual perceptions; or that what is awry with autistic children is that they lack a proper theory of mind.

References

G. P. Baker and P. M. S. Hacker (1985) *Wittgenstein: Rules, Grammar, and Necessity*, 1st edn (Oxford: Blackwell).

—— (2009) *Wittgenstein: Rules, Grammar, and Necessity*, 2nd edn (Oxford: Wiley-Blackwell)

G. Frege (1979) *Posthumous Writings* (Oxford: Blackwell).

P. M. S. Hacker (1986) *Insight and Illusion: Themes in the Philosophy of Wittgenstein* (Oxford: Clarendon Press).

P. M. S. Hacker (1996) *Wittgenstein: Mind and Will, Part I – Essays* (Oxford: Blackwell).

C. Koch (2004) *The Quest for Consciousness – A Neurobiological Approach* (Englewood, CO: Roberts and Company).

W. Köhler (1929) *Gestalt Psychology* (New York: Liveright).

1

Psychology's Inescapable Need for Conceptual Clarification

Daniel D. Hutto

> The confusion and barrenness of psychology is not to be explained by calling it a 'young science'; its state is not comparable with that of physics, for instance, in its beginnings…
>
> For in psychology there are experimental methods and conceptual confusion…The existence of the experimental method makes us think we have a means of solving the problems which trouble us, though the problems and the method pass each other by.
>
> Wittgenstein, *Philosophical Investigations* Sec. II, p. 232e[1]

1.1 Introduction

Wittgenstein offers a grave assessment of the state of psychology – one that falls just short of complete condemnation. Taken seriously, it should be a cause of concern for anyone working in the discipline today. But, should it been taken seriously? Was Wittgenstein's evaluation ever justified? More urgently, is it still an accurate portrayal of psychology as practiced today? This chapter argues it was and still is, and that this fact highlights an urgent and inescapable need for conceptual clarification in psychology. As a prelude to making this case, it is useful to get clearer about what motivated Wittgenstein's characterization of psychology as 'barren' because conceptually confused.

1.2 Wittgensteinian backstory

It might be thought that Wittgenstein's remark about psychology is inspired by nothing other than a general shunning or condemnation of

science, an expression of a general anti-scientific attitude. If so, perhaps his negative assessment of the condition of psychology can be put down to nothing other than an unjustified general dislike of science. While it might be tempting to assume this, such a reading lacks credibility.

The first thing to note is that when Wittgenstein speaks of conceptual confusions in the cited passage, he takes these to be of a distinctively philosophical kind – viz. they have a source and character that makes it impossible to overcome them by the provision of better or more refined theories, explanations or empirical studies. Focusing on this, noticeably, he makes no complaint about psychology's methodology, only its conceptual grasp of its subject matter. What this brings out is that the conceptual confusion that troubles psychology is really – at root – philosophical confusion (albeit exhibited in this instance by psychologists and not professional philosophers). Not surprisingly then, Wittgenstein reserves his most critical assessments not for the sciences but for the bankrupt state of philosophy and philosophers. It is philosophers who utter nonsense, who are in the grip of pictures, who are 'like savages, primitive people, who hear the expressions of civilized men, put a false interpretation on them, and then draw the queerest conclusions from it' (PI, §194). Hence, 'a philosophical problem has the form: "I don't know my way about"' (PI, §123). Scientists, however, have no natural immunity to intellectual diseases of this sort; and when they do suffer from them, the only possible treatment for their condition takes the form of philosophical work.

It is simply false that Wittgenstein is anti-science. What he does reject is the seductive scientistic view that the scientific method and outlook should dominate *all* thinking; that all problems and questions reduce to scientific ones. In particular, if the scientific method is not appropriate for dealing with philosophical problems of the sort that arise from conceptual confusions, then making new discoveries or acquiring deeper knowledge of phenomena through the manufacture of better, more penetrating theories cannot possibly solve or address them.

With this in mind, Wittgenstein rejects all attempts to provide philosophical explanations (which he thinks only breed more myths and confusions) in favour of a purely descriptive approach. He is utterly clear about the notion of explanation that he has in his sights and which he rejects:

> I mean the method of reducing the explanation of natural phenomena to the smallest possible number of primitive natural laws; and in mathematics, of unifying the treatment of different topics by using generalization.

> Philosophers constantly see the method of science before their eyes… I want to say here that it can never be our job to reduce anything to anything, or to explain anything. Philosophy really *is* 'purely descriptive.' (BBB, p. 18, emphasis original)

For Wittgenstein, conceptual investigations do not add to our understanding of what there is by positing new and unanticipated entities in the way that theorizing in a basic science like physics might.[2] For him, philosophical work does not add to our compendium of knowledge – it yields neither scientific knowledge nor a special kind of non-scientific knowledge. It elucidates and clarifies. By putting to bed – or at least temporally controlling – certain misleading habits of thought, it brings to light or reminds us of our genuine conceptual commitments, those that are integral to our everyday practices. This is achieved, on a case-by-case basis, by appeal to examples that reveal the point, context and specific character of our familiar modes of thought and talk. The positive part of philosophical work is the assembly of reminders about the contexts in which we actually deploy our concepts, the roles such concepts play in our practices and the point of such practices. But this kind of work is only possible if we can expose, and free ourselves from, tempting but misleading pictures that systematically block our view of how we actually use our concepts, something that lies before us in plain view, if we only have the eyes and will to see it.

In toto, philosophical labour of this sort should enable us to appropriately characterize and understand our quotidian commitments. This is the aim of clarificatory philosophy. Harking back to an early voiced, but constant thought of Wittgenstein's: 'Philosophy is not a body of doctrine but an activity' (TLP, 4.112). This is to realize – as Wittgenstein did throughout his career – that 'Philosophy is not one of the natural sciences. (The word "philosophy" must mean something whose place is above or below the natural sciences, not beside them)' (TLP, 4.111). It is against this backdrop that we can appreciate how and why Wittgenstein sums up his attitude to science in the following quotation:

> I may find scientific questions interesting, but they never really grip me. Only conceptual and aesthetic questions do that. At bottom I am indifferent to the solution of scientific problems; but not the other sort. (CV, p. 79e)

What we can take from all of this – the point that needs emphasizing – is that insofar as psychology is subject to conceptual confusions, these are

unavoidably philosophical in nature. That is why Wittgenstein makes no critique of the psychologists' experimental method of data collection. His concern targets how psychologists characterize and understand the data, not how they come by them.

1.3 Options for psychology

1.3.1 Acceptance

Total acceptance is one possible way for psychology to respond to the Wittgensteinian challenge. It might adopt the austere strategy of refusing to say anything that goes beyond the evidence. By sticking only to the collection of raw data provided through operationalized and controlled experimentation – by steering clear of all conceptual interpretations or speculation about such data – psychology can avoid the charge that it is conceptually confused by making sure it is conceptually empty. As long as psychology makes no attempt to understand or even characterize its subject matter, it protects itself against confusion and the need for conceptual investigation. Of course, with greater or lesser commitment to this positivist programme, radical behaviourists have already tried (and failed) to advance the fortunes of psychology by avoiding all conceptual and constitutive questions of a sort that could not be addressed through straightforward empirical experimentation.

Many scientists sport attitudes that appear to welcome this sort of approach; they are typically uninterested in, and impatient with, conceptual, constitutive questions. Burge (2010) provides the standard analytic philosopher's answer about the nature of such questions (which, when compared with the observations of the preceding section, shows itself to be highly un-Wittgensteinian), and he also identifies why most scientists adopt a negative attitude toward them.[3]

> A constitutive question concerns conditions on something's being what it is, in the most basic way. Something cannot fail to be what it is, in this way, and be that something. Constitutive conditions are necessary or sufficient conditions for something's being what it is in this basic way. To be constitutive, the conditions must be capable of grounding ideal explanations of something's nature, or basic way of being…*Science is more interested in finding explanations of how and why things happen than in asking about natures…Often good scientific work can proceed without answering constitutive questions correctly*. Still, obtaining clarity about key concepts, and delimiting boundaries of fundamental kinds indicated by such concepts, can strengthen and

> point scientific theory. It can help deepen understanding of frameworks within which scientific explanations operate. (Burge, 2010, p. xv, emphasis added)

Reliance on everyday mental concepts is utterly pervasive in psychology. It is part and parcel of the interpretation of data whenever and wherever everyday psychological phenomena are under investigation. This includes investigation of such familiar and core psychological phenomena as 'believing', 'experiencing', 'pretending', and 'empathizing'. Consider, for example, the following passage from a recent introduction of a special issue on empathy. It nicely illustrates – and gives the flavour of – a characteristic and undeniable fact about the state of the understanding of the concepts that frame and underpin psychological research. For despite employing robust experimental methods, it is admitted that:

> Almost anybody writing in the field would declare that there is no accepted standard definition of empathy – either among the sciences or the humanities or in the specific disciplines. (Engelen and Röttger-Rössler, 2012, p. 3)

Tellingly, despite initially recognizing that there is 'no established concept' at work, the authors of this introduction proceed to try to establish some working boundaries, persuaded by the essentialist assumption that there must be something – something with shared features at least, if not a single common feature – that makes it true that any and all cases of empathy are cases of empathy. And this sets the stage for what looks like an attempt at a traditional form of conceptual analysis designed to advance our understanding of the nature of empathy by determining to what extent, and in what measure, it involves 'thinking' and 'feeling', appealing to yet more everyday psychological concepts.

Of course, the promised deeper analysis in terms of these terms will turn out to be blunted if the concepts of 'thinking' and 'feeling' themselves turn out to lack 'agreed definitions', which, of course, they do.

Putting a proposal of the neuropsychologist Walter (2012) in the spotlight, the editors bring out the problems that attend a different sort of procedure that is neutral with respect to unpacking our conceptions of thinking and feeling. For in Walter's work:

> Empathy … is defined only as the understanding of the emotional state of the other and not by whether the process of understanding

> the emotional state is either an affective or a cognitive one. If it is a cognitive one, it is called cognitive empathy or affective theory of mind; if it is an affective one, it is called affective empathy. (Engelen and Röttger-Rössler, 2012, p. 5)

But here, too, there is no way to avoid reliance on everyday notions. For what understanding of understanding is in play? Here again, psychologists are prepared to acknowledge the serious difficulties in articulating exactly what it means to say that a child – or, indeed, anyone – has an understanding of some or other concept. As Apperly (2011) observes:

> It is common to ask when children understand perceptions, knowledge and belief *as such*. When can they distinguish them from other kinds of things and from each other? When do they understand how they come about and how they interact with each other? … *there is no easy answer to these questions*. *It is just not clear* whether we should credit a child a concept of 'knowledge' (for example) when she shows sensitivity to her mum's 'experience' at 14 months … when she first makes correct verbal judgements about someone else's lack of knowledge at 3 years, or when she first understands Oedipus problems at 6 years. (Apperly, 2011, p. 18, emphasis added)

Naturally, as scientists, who trust in the rigour and robustness of their experimental methods, psychologists will want to reject the idea that advances in their field – in any way – depend upon, or must wait for, philosophers to sort out conceptual matters by providing answers to constitutive questions. That would be an especially bitter pill to swallow, given that when such questions have been investigated by philosophers using the tools of conceptual analysis – proposal testing by means of trial by counterexample and intuition – they have rarely, if ever, identified the definitive necessary and sufficient conditions that constitute any phenomenon of interest.

Moreover, there is no doubt that good scientific work can proceed in the form of austere data collection without addressing constitutive questions or answering them correctly. But in so far as a science (in this case, psychology) is concerned to do something more than collecting raw data for possible interpretation – to the extent that it aspires to provide insights into the human condition – it cannot forever shun conceptual questions.

Conceptual questions are crucial not only in deciding which experiments are worth conducting and even framing how they might be best

conducted. And, of course, it would be impossible to draw meaningful conclusions from experiments without addressing such questions. Ultimately, not to do so makes it impossible to pass reliable judgment on the scope and nature of psychological investigations. For all of these reasons, the tactic of acceptance and avoidance looks like a non-starter.

1.3.2 Denial

If not avoidance by acceptance, then perhaps outright denial might work. Can it not be assumed that the phenomena somehow speak for themselves, obviating any need to investigate conceptual frameworks? If so, won't psychologists be able to get all they need for free? This is, of course, a naïve response to the problem and a quite hopeless one, given that it makes no sense to suppose that there could be pure, conceptually unmediated ways of getting at and investigating any subject matters of interest comprehendingly. Anticipating this type of reply, Wittgenstein notes and criticizes this common temptation:

> We feel as if we had to penetrate phenomena: our investigation however is directed not towards phenomena, but, as one might say, towards the 'possibilities' of phenomena. We remind ourselves, that is to say, of the kind of statement that we make about phenomena. (PI, §90)

To say this is not, of course, to endorse any kind of unwarranted or unworkable idealism or anti-realism. It is simply to note that it is impossible to conduct investigations into any subject without attending to *the ways in which we think about and conceive of* the topics under investigation. Putnam (1992) usefully clarifies the situation:

> To deny, as I do, that there is a 'ready-made world' is not to say that we make up the world. I am not denying that there are geological facts which we did not make up. But I have long argued that to ask which facts are mind dependent in the sense that nothing about them reflects our conceptual choices and which are 'contributed by us' is to commit a 'fallacy of division'. What we say about the world reflects our conceptual choices and interests, but its truth or falsity is not simply determined by our conceptual choices and our interests ... One thing that interests me ... is why we are so reluctant to admit this. What does it show about our culture and our entire way of thinking that it is so hard to admit this? (Putnam, 1992, p. 59, emphasis added)

It is for this reason that when we encounter philosophical difficulties and conceptual confusions, 'It is not the phenomenon under scrutiny that puzzles us here, although it seems to be; it is the language we employ to describe it' (Hyman, 1991, p. 6). Enlightened psychologists recognize that simply taking certain interpretations for granted – that is those that strike one as obvious – may not be an innocent reading of what is already and simply there. Thus, for example, in criticizing this habit as it figures in social cognition research, Apperly (2011) observes that:

> If we start out with an interest in mindreading there is a tendency *to see a need for it* in almost any social activity. *How else* could we explain an infant's ability to engage in a teasing interaction with its carer, a child's ability to understand everyday social interactions, or adults' remarkable ability to work out what one another are talking about? Surely in all these cases it is *necessary* to think about what other people know, think, want or intend? Actually, there are many reasons for thinking that this is often unnecessary. (Apperly, 2011, p. 114, emphasis added)

Others, too, recognize this problem, taking it a step further and putting it down to the influence of a particular kind of conceptual framework – a 'Theory of Mind' or ToM. Thus:

> the problem of behavioural data interpretation in ToM research studies is a special one: the researcher herself, in interpreting behavioural data in terms of children's possession of ToM, is in fact attributing mental states to others. In other words, *there is an intrinsic possibility of the researcher's interpretational activity being biased by her own possession of a ToM*. Such a possibility makes ToM studies dangerously susceptible to vicious circularity. (Dolcini, 2010, p. 42, emphasis added)

While this observation is intended to highlight something that psychologists ought to be wary of – that is the undue influence of their ToM – it also at the same time suggests another possible solution to the worry we have been examining. For can we not suppose that the conceptual problems of the sort that Wittgenstein highlighted can be resolved, at least in principle, by interrogating and making explicit the content of our everyday, built-in 'Theory of Mind'? Any problems that psychologists might face in answering constitutive questions about the nature of mental states and their defining characteristics should be addressable,

just in case it is safe to assume the existence of a ToM that incorporates well-defined principles.

1.3.3 ToM to the rescue – take one

In line with the acceptance of commonsense functionalism, many orthodox analytic philosophers of mind hold that (1) our everyday thought and talk about the mental is the expression of an implicit folk theory; (2) the meaning of our everyday psychological terms is fixed by the content of that theory; and (3) the content of that theory is characterized by what everyone finds intuitively obvious about the mind.

Wouldn't endorsing this kind of functionalism fit the bill very nicely for solving psychology's problem? Doing so provides a way of skirting the very worries that Wittgenstein raises about it. Indeed, when Hacker (this volume) traces the origins of Wittgenstein's remark about the confusions and barrenness of psychology, he reveals its target to be a claim made by Köhler. Specifically, Wittgenstein objected to Köhler's thought that, although the psychology of his day suffered from a lack of detailed knowledge of functional relationships of psychological phenomena, a mature science of psychology would emerge when precise functional laws were discovered. Once fully developed, psychology could iron out its conceptual difficulties if psychological phenomena could be functionalized. So conceived, psychology would be a special science, analogous to physics in core respects, but operating on a quite distinct, autonomous level.

Putnam (1967) regarded his brand of functionalism as just such a framework – as a 'schemata for hypotheses'. Hence, he hoped to find abstract functional laws for the behaviour of various psychological phenomena that would 'bring in its wake a delineation of the kind of functional organization that is *necessary and sufficient* for a given psychological state, as well as a *precise definition* of the notion "psychological state"' (Putnam 1967/2008, p. 45, emphasis added).

Putnam's own brand of functionalism is not ideal for addressing Wittgenstein's worries about psychology, because its abstract functional laws are too austere to capture all of the important features of our everyday mental state concepts. In particular, a pure, abstract functionalism is not well placed to account for the intentional or contentful properties of mental states. But if a built-in ToM exists – one that is universal to the human species – then it could provide the requisite defining structure and content that would secure our everyday psychological framework and – potentially – justify psychology's reliance on it.

Prima facie, this may look like the best chance of answering Wittgenstein's scepticism about psychology, providing a means for it to escape its conceptual confusions. But any psychologists who lean on the matter-of-fact existence of ToMs to provide a conceptual basis for their scientific endeavours face a dilemma. For either they must take the idea that a ToM exists to be (1) a substantive empirical hypothesis to be justified by evidence, or else it must be (2) an indisputable philosophical observation justified by other means. It turns out, on examination, that neither of these options provides psychology the solution it needs. Let us consider what happens if we suppose that the existence of a ToM has the status of (1) first. Psychologists who take this option are resting their futures on the shaky empirical bet that ToMs actually exist. But to the extent that the issue is empirically resolvable at all, this is far from a safe bet.

To see why, it is necessary to get clear about how the empirical issue might be decided. Matters here are not straightforward. Of late, concerns have arisen about whether the ToM hypothesis is empirically testable at all. As things stand, given the notion of a ToM that is in question, it is not obvious how to decide experimentally whether ToMs exist or not. Notably, it is not possible to use the methods of developmental or cognitive psychology in order to test directly whether a ToM exists. As long as a ToM is understood in the relevant sense – that is as a set of represented laws that define our core mental concepts – the best chance of moving forward would be to find some way of deciding which of the two rival theories, pure Theory Theory (TT) or pure Simulation Theory (ST), is true.

These theories logically exclude one another. Consider that if ST is true, then there are, as a matter-of-fact, no represented laws of folk psychology. The crucial difference between TT and ST is that the latter denies, while the former asserts, that 'the laws' that TT imagines are needed to explain our core folk psychological competence are subpersonally represented. ST insists that mindreading only requires using one's own cognitive machinery to model the mental states of others, obviating the need to represent the laws of folk psychology, at all.

So, to decide empirically whether ToMs exist boils down to testing which of these two theories is correct. But it is not at all clear that this can be done. This fact is becoming apparent to prominent researchers in the field. Thus, we are told:

> the debate between theory-theory and simulation-theory ... has been *remarkably poor at generating empirically testable hypotheses* for

> propositional mental states such as seeing, knowing and believing ... nobody has come up with a generalizable test, or a set of criteria, that could be used to discern whether a particular mindreading problem was solved by simulation or theory ... I am not the first, and probably not the last, to wonder whether experimental psychology might not be better off without these theories. (Apperly, 2011, p. 5, emphasis added)

There are systematic reasons why it is so hard to put the existence of ToMs to the empirical test. This is transparent if we remind ourselves what is really at stake in the debate between TT and ST. The reason is that the issue turns on whether the laws of folk psychology are internally represented, and this is not something that is directly empirically testable using purely psychological methods. This is why the existence of ToMs, understood in the relevant way, is not under active investigation *by psychologists*.[4]

Noting this is important, for once it is brought to light the natural move is for psychologists to pass the buck. Empirically establishing the existence (or otherwise) of ToMs is 'somebody else's problem'. Whether ToMs exist is *empirically* resolvable just in case the 'somebody' in question is *some other scientist*. But why assume that? After all, the idea that the TT–ST debate *must* be *empirically* resolvable is inspired by little more than the dogma that *any* properly formulated question about 'what there is' is a question that can be decided entirely by means of scientific investigation.

But is there another science that can decide the issue? Can we look to neuroscience, for example, to settle the matter? Eliminativist doomsayers have long cast negative predictions about the future of folk psychology, holding that it will be shown to be out of step with growing modern science. Ramsey highlights the key worry, telling us, 'Folk psychology is committed to the existence of mental representations. Therefore, for folk psychology to be vindicated, the correct scientific theory needs to invoke, at the very least, inner cognitive representations' (Ramsey, 2007, p. 114).

But many now doubt that such vindication is on the cards. Punchily, Churchland tells us that 'neuroscience is unlikely to find "sentences in the head" ... [and] on the strength of this assumption, I am willing to infer that folk psychology is false' (Churchland, 1991, p. 65). But philosophers and cognitive scientists involved in the general debate about the existence of mental representations are right to respond by pointing out that the issue cannot be settled by such direct appeals to neuroscience.

Whether or not mental representations exist must be decided at a higher level of abstraction than looking directly at the brain since neural activity, at best, only implements cognitive activity.

This observation is surely right, but it does not help those hoping to put the eliminativist charge to rest. This is because new developments in the *cognitive* sciences also provide compelling grounds to doubt the existence of mental representations; of the sort that folk psychology allegedly posits and those upon which the existence of ToMs depends. In sum, the anti-representationalist turn in cognitive science casts doubt on the existence of representations of precisely the kind that are needed in order for ToMs to exist (Chemero, 2009; Hutto and Myin, 2013; Ramsey, 2007).

Two things need to be kept separate at this juncture. The fate of the representational theory of mind is linked to the fate of folk psychology and the fate of ToMs in quite different ways. Arguably, despite all the hype, the future of folk psychology in no way depends on whether mental representations are vindicated by the cognitive sciences. In thinking otherwise, eliminativists commit the *ignoratio elenchi* fallacy; their prediction about folk psychology's future 'misses the target' because they misrepresent folk psychology's commitments; commitments which they wrongly take to be transparent and easily known. Those who think that folk psychology is due for elimination *because* mature cognitive science is likely to abandon mental representationalism have seriously misunderstood the commitments of the folk. Eliminativist assessments about the fate of folk psychology are really just negative bets on the likelihood of incorporating a *certain picture of folk psychology* into the scientific worldview. The picture of folk psychology in question takes the folk to be wedded to a particular (and very likely false) understanding of the nature of contentful mental states and the way such mental states cause actions. Specifically, eliminativists assume that the folk assume that mental states are representational and causally productive of behaviour in a way that directly competes with the sorts of impersonal explanations offered in the cognitive sciences and neuroscience. Although widely accepted, this is an unwarranted interpretation of the commitments of practicing folk psychologists.

Eliminativist predictions of a dim future for folk psychology rest on the assumption that it is, in essence, a kind of low-level theory – one that explains and predicts behaviour by positing the existence of propositional attitudes in the form of brainbound, contentful, mental representations that are casually efficacious in exactly the same way physical phenomena (broadly construed) are causal. Yet, close scrutiny

of what the folk are committed to in their interpretative, attributional practices – that is as revealed by the way the folk use their concepts when making sense of reasons – shows that this standard interpretation is without support (see Hutto, 2011). Once this is recognized, it becomes clear that the fate of folk psychology in no way depends on the outcome of debates about mental representations in the cognitive sciences.

But even if folk psychology might be safe if representationalism turns out to be false, the same cannot be said for ToMs. Their existence would be ruled out because ToMs are conceived of as cognitive devices that literally contain contentfully represented theories. After all, a ToM is imagined to be nothing but the set of mental representations of the laws of folk psychology. Thus, if the eliminativist predictions about representationalism turn out to be correct, then the existence of ToMs will not be vindicated by cognitive science in the long run. And, as it happens, there is every reason to think the eliminativists are right about the fate of representationalism (see Hutto and Myin, 2013).

What the twists and turns of the foregoing discussion are designed to highlight is that it is a bad strategy for psychologists to rest their discipline's future on the assumption that its conceptual framework can be secured by the existence of ToMs. This is to take a highly risky empirical bet. Psychologists would be betting their conceptual house on the outcome of debates about the existence of mental representations. That would require admitting that there is a very live possibility that there might be no conceptual basis for psychology as a distinctive special science if the evidence goes against them. Things will go this way if it happens that mental representations of the sort ToMs require happen not to exist. Consequently, those who adopt the strategy of taking an empirical bet on the existence of ToMs to secure the conceptual basis of psychology put its autonomy on the line, leaving its fate to the outcome of on-going theoretical debates elsewhere in the cognitive sciences.

Relying on the existence of ToMs in this empirical way seems too risky. If psychology wants to secure its conceptual framework in a more stable fashion, it will need to eliminate all risk in its portfolio. This might be achieved by turning to philosophers, not scientists, for aid. Although those who are purely empirically minded will naturally resist this, it seems a necessary move to achieve the desired end.

1.3.4 ToM to the rescue – take two

Analytic Functionalists – David Lewis (1970, 1972), Frank Jackson (1998) – assume that folk psychology is an implicit, term-introducing

theory, the core content of which can be captured by analysing folk platitudes about the mental. Such platitudes are assumed to capture the theory's content by revealing what the folk find obvious about the mental. So, in what sense is folk psychology an implicit theory? Lewis (1994) says that folk psychology is a body of tacit knowledge rather like our knowledge of syntax. Jackson elaborates, telling us that the idea that folk psychology is a theory depends on 'the availability ... of sentences that capture what [the folk] believe – of, that is, sentences that *represent as their minds do* when they believe that P, where P is the theory we are talking about' (Jackson, Mason, and Stich, 2009, p. 59, emphasis added).

But ToMists cannot take Lewis and Jackson at their words if they are to avoid positing the existence of ToMs as a mere empirical hypothesis. For if we are entitled to know, without any risk, that folk psychology is a theory, then the fact that it is, had better not imply the truth of the representational theory of mind (RTM), where RTM takes it that minds contain truth-evaluable representational contents capable of causally driving thought and action. The trouble, as we just saw, is that RTM is a substantive, empirical proposal about the nature of minds – one that may well turn out to be false.

If that commitment is in play, the only way that the existence of ToMs could be guaranteed in advance would be if it were also guaranteed in advance that RTM is true. But, illustratively, to rule out in advance the *possible falsity* of RTM would beg the question against traditional eliminativism. And this is a question that Analytic Functionalists intend to leave open. Hence, the most natural reading of Lewis's and Jackson's remarks about the status of folk psychology being an implicit theory must be rejected.

A lighter reading is possible. For Jackson also tells us that 'to have a theory is to have a certain view about how things are' (Jackson, Mason, and Stich, 2009, p. 87). Since the folk make use of mental concepts in making sense of others, it is safe to say that they have a view about how things are, with respect to the mental life of people. For Jackson, this equates to having a certain view about how things are – hence a theory. Note that this is not to say what makes it possible for the folk to have such a theory; hence, there need be no necessary commitment to RTM. But if one goes for this light reading, it is possible to regard the truth of theory-theory as 'near enough analytic' (Jackson, 1999, p. 80).

Read this way, psychology gets its desired result – there is no risk in the conceptual portfolio. Folk psychology names a contentful and structurally well-defined theory implicitly held by the folk. The cost is that

having a ToM is compatible with *any* empirical proposal about the basis of folk psychology – including Modular TT, Scientific TT, Simulation Theory and the Narrative Practice Hypothesis.

While this might solve the problem psychology faces, there are reasons for thinking that the story proposed by Analytic Functionalism is too incredible to be believed. One major problem relates to the questions about how we are to decide which platitudes express the folk theory. Analytic Functionalism or TT assumes it is possible to accurately characterize folk psychology's commitments in a logically perspicuous way. The method proposed is as follows:

> Collect all the platitudes ... *regarding the causal relations* of mental states, sensory stimuli, and motor responses. ... Add also all the platitudes to the effect that one mental state falls under another ... Perhaps there are platitudes of other forms as well. Include only the platitudes which are common knowledge amongst us: everyone knows them, everyone knows that everyone else knows them, and so on. (Lewis, 1972, p. 256, emphasis added)

Once the platitudes are in hand, it is possible to express the theory in a single sentence (a Ramsey sentence) which articulates the 'postulate' that is folk psychology, revealing the distinctive roles played by the various mental state concepts and how they stand in relation to other things. But this assumes that old-school conceptual analysis is in good working order. It assumes that it is possible to get at the content of our folk theory by 'appeal to what seems most obvious and central [about the domain in question], as revealed by our intuitions about possible cases' (Jackson, 1998, p 31).[5]

The proposed strategy for discerning the true commitments of the folk depends on discovering what everyone takes to be 'obvious'. So, again, how do we know that some platitude must be part of the folk theory? Answer: 'Because so many philosophers find it so very obvious. I think it seems obvious because it is built into folk psychology. Others will think it is built into folk psychology because it is so obvious; but either way, the obviousness and the folk psychological status go together' (Lewis, 1999, p. 328).

Weinberg et al. (2006) dub this approach Intuition Driven Romanticism. For what if another philosopher has contrary intuitions, as they often do? How, then, would we decide what the content of the folk theory really is? Here we must note that when such disputes arise, as they inevitably do, it is no good for those on one or the other side of the

matter to insist that they alone are really appealing to what is obvious. Indeed, 'to *insist* on the obviousness of anything is self subverting since the need for insistence contradicts the claim to self evidence and positively invites an opposing insistence' (Mulhall, 2007, p. 9, emphasis added).[6]

A more objective method is needed. Enter experimental philosophy! X-phi, as it is also known, seeks to overcome this problem for conceptual analysis by appeal to the scientific methods of the empirical sciences. Experimental philosophers run systematic experiments in order to reveal what ordinary people actually think about a given domain – for example. the mental and the moral (Knobe, 2007; Knobe and Nichols, 2008). Contra Lewis, the basic assumption behind such efforts, which is surely right, is that what philosophers find intuitive and obvious will diverge from what the folk find intuitive and obvious. Thus, deciding what folk psychology 'says' is, in the end, an empirical matter (Stoljar, 2009, p. 126).

But the truth is, as anyone who has tried to do this empirical work soon discovers, it is riddled with methodological problems (Kauppinen, 2007; Cullen, 2010). Moreover, the commitments of folk psychology cannot be 'easily extracted from the kind of things people say' (Ratcliffe, 2007, p. 49). Nor do they come in the form of a cluster of neatly interlaced principles, as Analytic Functionalists assume they will.

Recent times have seen more general suspicions aired about the role that intuitions are meant to play in all of this.

> 'Intuition' plays a major role in contemporary analytic philosophy's self-understanding. Yet there is no agreed or even popular account of how intuition works, no accepted explanation of the hoped-for correlation between our having an intuition that P and its being true that P. Since analytic philosophy prides itself on its rigor, this blank space in its foundations looks like a methodological scandal. Why should intuitions have any authority over the philosophical domain? (Williamson, 2007, p. 215)

Building on this, more fundamentally, we must seriously question whether what the folk find obvious or intuitive necessarily expresses the content of implicit folk theories. For even if all of the aforementioned problems could be overcome, there is always the risk that the intuitions in question will be shaped by distorting pictures – pictures that have nothing to do with the way our concepts are used in our everyday practices. To probe what the folk find obvious when they think about

the mental may not tap a philosophically uncontaminated source of understanding. There is a clear and present danger that prominent folk intuitions about a given subject matter will be informed, not by what is integral to folk practices themselves, but by, say, certain popular pictures of the nature of mind – even if only indirectly.

The deeper worry – which concerns the most serious problem in relying on intuitions, whether in old-school or new-fangled ways – is that doing so runs the risk of allowing philosophical commitments to taint the outcome of any conceptual analysis. If so, we will only ever get a philosophically influenced view of what is obvious about some domain, not the real folk view. This is not easy to notice, for, as Hyman stresses:

> Pictures ... wreak havoc if, when we theorize about the mind, we take them at face value ... Their grip on the imagination is not the grip of a tremendous hypothesis, like the big bang, but more like the grip of an entrancing metaphor or myth; and their influence on theory is as permanent as the language in which they are lodged. (Hyman, 1991, p. 7)

Worse, as Wittgenstein underscored, they are hidden in plain view, for 'when we have got a picture of our ordinary way of speaking we are tempted to say that our way of speaking does not describe the facts as they really are' (PI, §402). Putting all of this together:

> We need to be wary of the intuitions we have when doing philosophy. Such intuitions are frequently accorded some sort of evidentiary value; the intuition p is often taken to support the assumption that p. Indeed intuitions have been traditionally regarded as immediate insights ... [but] philosophical intuitions need not result from the exercise of a competence, they may be due to cognitive distortions. (Fischer, 2011, p. 52)

The real worry is that:

> Philosophical intuitions are not implied by ordinary language statements. They are not pieces of common sense as implicit in ordinary language. They cannot be put down to the mere exercise of ordinary linguistic, recognitional or classificatory skills. (Fischer, 2011, p. 33)

What might it mean for intuitions to be picture-inspired? Here we can call on psychology to do interesting work for philosophy. For, according to Fischer,

> A thinker is *under the spell of*, or adheres to, *a philosophical picture* iff he systematically makes non-intentional analogical inferences which assimilate targets to models of a conceptual metaphor, in ways the thinker knows them to be different. (Fischer, 2011, p. 32, emphasis original)

Fischer's claim is that 'Non-intentional analogical inferences lead to conclusions that the thinkers who make them are prone to find intuitively compelling' (Fischer, 2011, p. 33). And once we start relying on intuitions produced in this way, 'musty' thinking takes over – we begin to think that we have understood how things must be in some domain, without argument, indeed without even the possibility of argument. Calling on psychology once again, 'Once picture-driven reasoning has led us to [a] philosophical conception, belief-bias effects may have us read this conception into ordinary talk' (Fischer, 2011, p. 45). Thus, on the basis of this sort of analysis, Fischer, like Wittgenstein before him, bids us to recognize that philosophical questions and problems that arise from picture-driven reasoning cannot be solved but only dissolved by means of systematic diagnostic analysis (Fischer, 2011, p. 77, see also p. 72).

Are there any tell-tale signs that our thinking about some topic has become *philosophically* contaminated by attachment to a deluded picture or ideology? Yes: that such thinking leads to intractable philosophical conundrums relating to the topic domain (e.g. the problem of mental content, the problem of mental causation). Also, those in the thrall of philosophical pictures are unable to articulate a clear, stable content – as evidenced by their vacillating between positive and negative sublimation – of only describing the *explanandum* (not the *explanans*), on the one hand, or merely saying what the theory does *not* claim, on the other – when trying to express the positive content of their philosophical theories.

In sum, even if a reliable way could be found for capturing 'shared' folk intuitions objectively, there is always the serious risk that the intuitions in question are rooted in commonly held, but distorting, pictures about the topic domain. If so, they will not reveal the content of an implicit and shared folk theory that underwrites or is otherwise embedded in folk psychological practice. Thus, such an unveiling would tell us nothing about how our concepts are used in everyday folk psychological practice. The root problem is, again, that probing what 'the folk' find obvious about the mental need not be to tap a philosophically pure and uncontaminated source.

Despite its seductive attractions, the Analytic Functionalist characterization – which is now the default, accepted framework for defining

the meaning of mental concepts in analytic philosophy of mind – is not supported by careful attention to what the folk do. That is hardly surprising, since folk psychology did not originate in philosophical reflection upon what people actually think. Instead, it arose through the imposition of a set of pre-formulated philosophical assumptions upon the "folk"' (Ratcliffe, 2009, p. 382).

1.4 A different way

All of these problems dissipate if one abandons an interest in revealing what the folk 'find obvious' in favour of trying to discover what is integral to their competent use of concepts, when these concepts are put to good work. Taking the case in hand, the first step in making this shift is to stop thinking of folk psychology as an implicit theory. With far less baggage, we can assume that folk psychology denotes – at a bare minimum – the everyday business of making sense of intentional actions (i.e. our own and those of others) in terms of reasons. It requires being able to answer a particular sort of 'why' question by competently deploying the idiom of mental predicates (e.g. beliefs, desires, hopes, fears). So conceived, folk psychology is 'how people actually understand each other's behaviour, rather than an account of how they think that they think in interpersonal scenarios. People may not have a clear idea about what is central to their thinking about others, and so [folk psychology] must be distinguished from what we might call "folk folk psychology"' (Ratcliffe, 2009, p. 381).

Folk psychology as a practice is a perfectly familiar, out-in-the-open, activity, not something hidden away in the minds of the folk. If folk psychology is understood in this way, then there is no point in trying to probe commonplace intuitions in order to reveal its core commitments. There is no need to try to discover the content of an implicit theory held by the folk, since we have exactly no reason for presuming that there is any such thing.

It is a fundamental error to suppose that our conceptual investigations ought to target (1) what the folk 'find obvious' about a given domain (which is putatively revelatory of a shared implicit theory) instead of (2) attending to what the folk do when competently deploying their concepts in dealing with that domain. Only the latter reveals the folk commitments of interest. Focusing on what the folk find obvious, as Analytic Functionalists or TTists claim to do, generates a host of methodological difficulties that are best avoided. Much worse than this, trying to identify what is 'intuitively known by all' typically results

in contaminated pictures, of the genuine commitments of the folk, hogging our attention.

The cardinal sin of Analytic Functionalism or TT is that it makes it appear as if it is a simple matter to obtain an accurate understanding of folk commitments. Focusing on what anyone and everyone will find 'obvious' about some domain aids, abets and seemingly legitimizes certain popular, but biased, pictures of our folk commitments. This becomes dangerous when, by fuelling our intuitions, such pictures set important philosophical agendas and play a leading role in evaluating the adequacy of philosophical proposals.

The sorts of conceptual investigations Wittgenstein engages in are nothing remotely like that of those who hope to reveal substantive philosophical truths based on the *a priori* analysis of concepts. We are brought to understand our practical commitments by being reminded both of the way we operate with concepts in specific contexts and of the point of doing so. There seems no good reason for denying that this is at least one of the things that philosophy can do – indeed, it is one of the good things that philosophy can do when it does its work well.

So conceived, the ambitions of this sort of philosophical activity are distinct from the ambitions of scientific investigations. Debunking certain misleading philosophical pictures of what we are committed to with respect to our everyday thought and talk about the mind. Thus:

> Investigating which concepts are fundamental to our thinking ... is not an easy task. The fruits of such an enquiry need to be distinguished from superficial and possibly widespread intuitions. (Ratcliffe, 2009, p. 380)

Indeed, the result of such an investigation, properly conducted, would be 'a debatable philosophical position [taking] considerable philosophical work to formulate, rather than a casual, uncontroversial statement of what the "folk" think' (Ratcliffe, 2009, p. 382). This is surely right, as revealed, for example, by the on-going debate between Ratcliffe and myself about whether, once one accepts this methodology, one finds there is no residue of 'folk psychology' left in our everyday conceptual practice (See Hutto, 2008b; Ratcliffe, 2008, 2009). Against Ratcliffe, I maintain there is no compelling grounds to accept that folk psychology 'is so abstract and uninformative that it warrants rejection rather than revision' (Ratcliffe, 2009, p. 386). Specifically, I reject the idea that no vestige of it forms part of the structure of everyday interpersonal understanding.[7] This is not the place, for reasons of space, to attempt to settle

that dispute; rather, I wish only to highlight that it is precisely the sort of fruitful debate that requires conceptual clarification of a kind that psychology inescapably needs.

Notes

1. I have used this quotation before to open and focus a chapter on this topic along similar lines (see Hutto, 2009). This chapter is a kind of 'take two'; a second attempt to explore the relevance of Wittgenstein's pivotally-important assessment of the state of psychology to current understanding of the discipline.
2. The approach is antithetical to that adopted by most mainstream philosophers who see the job of philosophy to provide theories that augment or advance our scientific understanding. Accordingly, in line with that view, 'one rather hopes that there will prove to be many more – and much odder – things in the mind than common sense had dreamed of; or else what's the fun in doing psychology?' (Fodor, 1987, p. 15).
3. While there is a way of reading Burge's (2010) remarks that is amendable to the philosophical approach I advocate in this chapter, it is clear on close scrutiny that Burge adopts an un-Wittgensteinian approach to conceptual investigations. This shines through when, for example, Burge observes that 'When done well, philosophy has made some impressive contributions toward clarifying basic concepts and reflecting on basic kinds invoked in the sciences. Such contributions are less infrequent, and tend to be more fundamental, with new and maturing sciences' (Burge, 2010, p. xv). In saying this, Burge is assuming that philosophy helps us to gain a deeper understanding of the phenomena in question by providing theories of the nature of things and, strikingly, if we were to apply the above thought to the case of psychology, we might well put its conceptual confusion down to its being a 'young science'; just the opposite of what Wittgenstein suggests.
4. The fact is that 'empirical research provides us with a sophisticated understanding of the developmental stages, the conceptual sophistication associated at each level, and the biological mechanisms implementing our folk psychological abilities. It should not be conceived as deciding the debate between theory theorists and simulation theorists' (Stueber, 2006, 152).
5. As Jackson puts it: 'My intuitions about possible cases reveal my theory. ... Likewise, your intuitions reveal your theory. To the extent our intuitions coincide with those of the folk, they reveal the folk theory' (Jackson, 1998, p. 32).
6. For a more detailed discussion of the worry about the objectivity of traditional forms of conceptual analysis and other related problems, see Hutto 2003/2006, ch. 6, Sec 3.
7. In a similar sceptical vein, Strijbos and de Bruin (2012) raise the possibility that the Belief-Desire model may not be, as originally intended, 'a proper, personal-level description of what people do when they interpret one another in terms of their reasons for action' (p. 145).

References

I. Apperly (2011) *Mindreaders: The Cognitive Basis of 'Theory of Mind'* (Sussex /New York: Psychology Press).
T. Burge (2010) *The Origins of Objectivity* (Oxford: Oxford University Press).
A. Chemero (2009) *Radical Embodied Cognitive Science* (Cambridge, MA: MIT Press).
P. M. Churchland (1991) 'Folk Psychology and the Explanation of Human Behaviour' in J. D. Greenwood (ed.) *The Future of Folk Psychology* (Cambridge: Cambridge University Press), 51–69.
P. Churchland (2007) 'The Evolving Fortunes of Eliminativist Materialism' in B. McLaughlin and J. Cohen (eds) *Contemporary Debates in Philosophy of Mind* (Oxford: Blackwell Publishing), 160–82.
S. Cullen (2010) 'Survey-Driven Romanticism', *Review of Philosophy and Psychology*, 1(2), 275–96.
N. Dolcini (2010) 'Minding the Developmental Gap: A Theoretical Analysis of Theory of Mind Data', *Journal of Consciousness Studies*, 17(7–8): 37–46
E.-M. Engelen and B. Röttger-Rössler (2012) 'Current Disciplinary and Interdisciplinary Debates on Empathy', *Emotion Review,* 4(1), 3–8.
E. Fischer (2011) *Philosophical Delusion and Its Therapy: Outline of a Philosophical Revolution* (London: Routledge).
J. A. Fodor (1987) *Psychosemantics* (Cambridge, MA: MIT Press).
D. D. Hutto (2003/2006) *Wittgenstein and the End of Philosophy: Neither Theory nor Therapy* (Basingstoke: Palgrave Macmillan).
D. D. Hutto (2008a) 'Limited Engagements and Narrative Extensions', *International Journal of Philosophical Studies*, 16, 419–44.
D. D. Hutto (2008b) 'The Narrative Practice Hypothesis: Clarifications and Implications', *Philosophical Explorations*, 11, 175–92.
D. D. Hutto (2009) 'Lessons from Wittgenstein: Elucidating Folk Psychology', *New Ideas in Psychology*, 27, 197–212.
D. D. Hutto (2011) 'Presumptuous Naturalism: A Cautionary Tale', *American Philosophical Quarterly*, 48(2), 129–45.
D. D. Hutto and E. Myin (2013) *Radicalizing Enactivism: Basic Minds without Content* (Cambridge, MA: MIT Press).
J. Hyman (1991) 'Introduction' in J. Hyman (ed.) *Investigating Psychology* (London: Routledge), 1–26.
F. Jackson (1998) *From Metaphysics to Ethics* (Oxford: Oxford University Press).
F. Jackson (1999) 'All That Can Be at Issue in the Theory-Theory Simulation Debate', *Philosophical Papers*, 28, 77–96.
F. Jackson, K. Mason, and S. Stich (2009) 'Folk Psychology and Tacit Theories: A Correspondence between Frank Jackson and Steve Stich and Kelby Mason' in D. Braddon-Mitchell and R. Nola (eds) *Conceptual Analysis and Philosophical Naturalism* (Cambridge, MA: MIT Press), 45–97.
A. Kauppinen (2007) 'The Rise and Fall of Experimental Philosophy', *Philosophical Explorations*, 10(2), 95–118.
J. Knobe (2007) 'Experimental Philosophy', *Philosophy Compass*, 2(1), 81–92.
J. Knobe and S. Nichols (2008) 'An Experimental Philosophy Manifesto' in J. Knobe and S. Nichols (eds) *Experimental Philosophy* (Oxford: Oxford University Press), 3–14.

D. K. Lewis (1970) 'How to Define Theoretical Terms', *Journal of Philosophy*, 67, 427–46.
D. K. Lewis (1972) 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy*, 50, 249–58.
D. K. Lewis (1994) 'Reduction of Mind' in S. Guttenplan (ed.) *A Companion to the Philosophy of Mind* (Oxford: Blackwell), 412–31.
D. K. Lewis (1999) *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press).
S. Mulhall (2007) 'Wittgenstein's Private Language: Grammar, Nonsense, and Imagination' in *Philosophical Investigations* (Oxford: Oxford University Press), §§ 243–315.
H. Putnam (1967/2008) 'The Nature of Mental States' reprinted in (2008) *Mind, Language and Reality: Philosophical Papers*, Vol. 2 (Cambridge: Cambridge University Press).
H. Putnam (1992) *Renewing Philosophy* (Cambridge, MA: Harvard University Press).
W. M. Ramsey (2007) *Representation Reconsidered* (Cambridge: Cambridge University Press).
M. Ratcliffe (2007) *Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation* (Basingstoke: Palgrave Macmillan).
M. Ratcliffe (2008) 'Farewell to Folk Psychology: A Response to Hutto', *International Journal of Philosophical Studies*, 16, 445–51.
M. Ratcliffe (2009) 'There are No Folk Psychological Narratives', *Journal of Consciousness Studies*, 16, 379–406.
D. Stoljar (2009) 'The Argument from Revelation' in D. Braddon-Mitchell and R. Nola (eds) *Conceptual Analysis and Philosophical Naturalism* (Cambridge, MA: MIT Press), 113–37.
K. R. Stueber (2006) *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences* (Cambridge, MA: MIT Press).
D. W. Strijbos and L. C. de Bruin (2012) 'Making Folk Psychology Explicit: The Relevance of Robert Brandom's Philosophy for the Debate on Social Cognition', *Philosophia*, 40, 139–63.
H. Walter (2012) 'Social Cognitive Neuroscience of Empathy: Concepts, Circuits, and Genes', *Emotion Review*, 4(1), 3–8.
J. M. Weinberg, S. Nichols and S. Stich (2006) *Philosophical Topics*, 29(1/2), Reprinted from 2001, 429–60.
T. Williamson (2007) *The Philosophy of Philosophy* (Oxford: Blackwell).

2

Wittgenstein's Method of Conceptual Investigation and Concept Formation in Psychology

Oskari Kuusela

2.1 Introduction

The goal of Wittgenstein's conceptual or grammatical investigations is the clarification of concepts and linguistic locutions we use to think and express our thoughts, with particular focus on their employment in formulating philosophical questions and answers. The purpose of this engagement with language is the resolution of philosophical problems connected with conceptual unclarities and confusions. Here the notion of a philosophical problem should not be construed too narrowly, however, as if these were problems for the philosopher only, and conceptual problems did not arise in connection with, for example, scientific thinking. As Wittgenstein remarks: 'A scientist says he pursues only empirical science or a mathematician only mathematics and not philosophy, – but he is subject to the temptations of language like everyone; he is in the same danger as everyone else and must beware of it' (MS 151, p. 6).[1]

Conceptual problems in Wittgenstein's sense pertain to our modes of expression and (re)presentation of reality, including our modes of (re)presenting or trying to clarify the uses of language for ourselves, for example, by means of definitions. One key characteristic of such problems is that they can get us stuck in our attempts to understand whatever we are trying to understand. In such a case, the modes of expression or (re)presentation we employ for the tasks of thinking at hand, including our construals of modes of expression taken over from extant domains of discourse, prevent us from rendering comprehensible our objects of investigation. (One example of concepts taken over from another domain would be the use of everyday psychological concepts, construed in particular ways, in scientific psychology.) In order to resolve such

problems, we need to reflect on and clarify relevant modes of expression or (re)presentation. New modes of expression and (re)presentation need to be devised in place of misleading ones. This, briefly, is what conceptual investigation does.[2]

As regards the relevance of Wittgenstein's work for psychology, two distinct albeit connected issues should be distinguished. On the one hand, it is Wittgenstein's view that conceptual confusion is widespread in psychology (PI II, xiv; RPP I, §1039). On his account, this is largely due to the assumption of simplistic accounts of the logical or grammatical role, function or use, of everyday psychological expressions as the starting point of psychological research. In this vein, Wittgenstein characterizes, for example, the idea of psychophysical parallelism that assumes mental phenomena to have straightforward neural correlates as 'a fruit of the primitive conception of grammar' (MS 134, p. 100/ RPP I, §906/Z, §611; I will return to this remark briefly at the end of the chapter). One way in which Wittgenstein may be regarded as relevant for psychology then is this: His investigations of particular psychological concepts and other relevant linguistic locutions can help psychological research to find its feet in the foggy field of the use of everyday psychological language, and thus help psychology to find fruitful perspectives and directions, whenever everyday notions are the starting point of psychology, as they often are. In particular, as Wittgenstein notes, conceptual problems cannot be solved by means of empirical research and experiments, but here the 'problem and method pass one another by' (PI II, p. xiv). Moreover, even when everyday concepts are rejected by psychology, it is still important to be clear about what is rejected, and not confuse the rejection of a philosophical (mis)construal of a concept with the rejection of an everyday concept.[3]

One the other hand, there are also more general questions about the status or role of concepts and principles in terms of which psychologists spell out their understanding of the nature of their objects of investigation, and that constitute the framework for empirical psychological research. How does, for example, a definition of 'thinking' employed in developing a psychological model for thinking or as the basis of experimental studies of thinking, relate to empirical, factual statements about thinking? What kind of a proposal is made when putting forward such a definition as the basis of empirical research? Furthermore, when everyday psychological concepts or some particular construals of them are used in psychology, how should the relation between everyday concepts and psychological concepts be understood? Is it an assumption or implication of Wittgenstein's approach to clarifying psychological concepts

by investigating their uses in everyday life that scientific psychology is bound by everyday concepts and should not seek to modify them? Should we expect there to be some kind of a correspondence between everyday psychological concepts and those of scientific psychology?

The focus of this chapter is on the latter kind of questions rather than issues pertaining specifically to any particular psychological concepts. I will discuss Wittgenstein's relevance for psychology from the perspective of his ideas concerning the methodology of conceptual investigation, his philosophy of science, or more specifically in the light of his notion of grammatical statements as statements of norms of expression, (re)presentation or explanation. In this way, I seek to clarify, firstly, the general import of Wittgenstein's clarifications of particular concepts, that is what kind of a statement is made when such a clarification is offered. Secondly, I address the question of whether certain methodological ideas of Wittgenstein's regarding conceptual investigation could find an analogous application in psychology, and thus help to resolve difficulties relating to the conceptualization of psychology's objects of investigation. Essentially, Wittgenstein's methods are designed for dealing with conceptual complexity, that is to enable us to approach dynamic and complex linguistic practices in a way that makes this complexity manageable. Perhaps it will ultimately be necessary for psychology, too, to acknowledge the non-reducible complexity of its objects of investigation, for example, that the phenomena of thinking, remembering or seeing do not constitute simple uniform unities but a multitude of varied cases falls under each concept, as Wittgenstein's investigations of psychological concepts suggest. In that case, perhaps Wittgenstein's methodological ideas can also help psychology to deal with complexity. Let us begin with an outline of Wittgenstein's method.

2.2 The method of conceptual investigation

Wittgenstein's conceptual investigations cannot be characterized simply by saying that their object of study is concepts or their linguistic expressions. Concepts can be the object of historical studies and metaphysical claims, too, but Wittgensteinian conceptual investigation differs from both. Instead, what is distinctive about his approach can be explained with reference to the status or role of grammatical statements by means of which concepts and the function of linguistic expressions can be clarified. Here it is crucial that such statements are not true/false in the sense of factual statements, including empirical statements about the uses of language and alleged 'super-factual' metaphysical statements about

necessities and possibilities pertaining to reality or thought or language use. Rather than stating something true/false, grammatical statements may be understood as expressions of grammatical rules, it being part of the concept of a rule that a statement of a rule does not state anything true/false about anything.[4]

The employment of grammatical statements or rules in philosophical clarification can now be explained as follows. Grammatical rules, Wittgenstein maintains, can be used to describe language use. (See MS 140, p. 15r/PG, p. 60.) Here the notion of description is to be understood in a particular sense, however, and it is important to observe a certain ambiguity pertaining to the word 'description'. On the one hand, as Wittgenstein explains, there are true/false descriptions of objects, such as houses or trees. In such a case, the object of description is given (or exists) independently of its description, which is also a condition for the possibility of it representing the object truly/falsely. This is the sense in which a factual statement describes something. On the other hand, there are also descriptions in the sense of designs and constructions, whereby what is described is not given (or does not exist) independently of the description. Accordingly, the description is not true/false about anything with which it could be compared for truth/falsity. Rather, the description depicts something *possible*. It is in this latter sense that statements of a rule can be used to describe a game, a calculus or the use of a word, but, as Wittgenstein emphasizes, such a description only describes a possible game, calculus or possible use for a word. A description by means of rules as such does not involve any claim or assumption about whether anyone ever actually played or will play such a game, or used or will use signs according to relevant rules. The point may be explained thus: Assume that someone spells out a rule for the use of a word, such as a definition. We may subsequently make use of the rule as a model for the functioning of a word, and thus employ it to describe the word's use. Clearly, however, to merely spell out a rule is not yet to compare reality, that is the actual use of any word, with the rule as a model for language use. The rule/definition might also be used for a number of different purposes, for example, stipulating a new use or ruling out a certain use. Hence, to represent actual use by means of the rule is a logically distinct further step from simply articulating a rule. (For discussion of the two senses of description, see MS 113, p. 28v/TS 211, p. 576/TS 213, p. 245r; cf. MS 115, p. 59.)

Still, philosophical clarification aims to resolve *actual* unclarities. To achieve this, conceptual investigation needs to render

perspicuous actual uses of language, not merely construct and speak about possible ones. Grammatical rules must therefore be brought into contact with actual language use. However, rather than involving any claims about the rules according to which language is actually used (in either an empirical or metaphysical sense of what those rules must be), Wittgenstein's later method consists in *comparing* actual uses with grammatical rules, whereby such a rule constitutes mode of (re)presenting the actual use of a word or some specific aspect of the word's use. Or, as he also puts it, clarification in terms of grammatical rules is a matter of describing language *in the form of a game* according to rules, without assuming or claiming that its users are actually playing a game according to any definite rules. (The point of avoiding claims about language use is elucidated below.) Wittgenstein explains this idea in the *Investigations* with the help of an example of '... people amusing themselves in a field by playing with a ball so as to start various existing games, but playing many without finishing them and in between throwing the ball aimlessly into the air, chasing one another with the ball and bombarding one another for a joke and so on. ...' (PI, §83) Clearly, there is no single definite game played here, and similarly Wittgenstein thinks language need not, and often is not, used according to any definite rules. Its users may switch between different definitions as they go along without this constituting a problem. (PI, §79–80) Still, according to Wittgenstein, the activities of the people in the example can be described as if they were playing a rule-governed game, and 'following definite rules at every throw' (PI, §83). As he explains in connection with an earlier version of the remark: '... while it is possible to give a rule for every action (move) which it corresponds to, we must in certain cases describe the use of language as a continuous change of the game (schedule of rules). ... So that we must say we view language *in the form* of a game, of acting according to a schedule of rules' (MS 112, pp. 95r–v; cf. TS 211, p. 492). In this sense, we can then describe uses of linguistic expressions by stating rules for their use, without claiming that they are actually used exactly in this way, and without having to insist that language users are playing a definite game according to rules. Rather, the rules are used to capture patterns of interest in actual use or aspects thereof for the purpose of clarification. The point can be further elucidated as follows.

Grammatical rules, or models constituted by such rules or systems thereof, are a means or an instrument for describing the function of linguistic expressions. By means of such rules, that is by comparing actual language use with them, specific aspects of actual language use

can be highlighted, captured and described, without making any claims about the possibility of reducing the actual complicated and dynamic uses of language to something simpler and static, such as use according to (a system of) definite grammatical rules. Wittgenstein explains the idea of describing specific aspects of dynamic uses in terms of static models as follows:

> If we look at the actual use of a word, what we see is something constantly fluctuating.
>
> In our investigations we set over against this fluctuation something more fixed, just as one paints a stationary picture of the constantly altering face of the landscape.
>
> When we study language we *envisage* it as a game with fixed rules. We compare it with, and measure it against, a game of that kind.
>
> If for our purposes we wish to regulate the use of a word by definite rules, then alongside its fluctuating use we set up a different use by codifying one of its characteristic aspects. (MS 140, p. 33/PG, p. 77, emphasis original)

Thus, although language use, according to Wittgenstein, is typically not governed by definite fixed rules, it can be presented for particular purposes of clarification as if it were used according to definite fixed rules. As he puts the point in the *Investigations*, this is to present '… the model is as what it is, an object of comparison …' (PI, §131; cf. §130). Such descriptions may be characterized as abstractions from actual more complicated uses or as idealizations, although, as Wittgenstein emphasizes, idealization is to be understood here in a specific sense. When speaking about language, philosophy does not speak about language in an ideal sense, as if its statements did not concern actual languages (or language as a spatio-temporal phenomenon; PI, §108), but something ideal that underlies actual languages, and which the latter only approximate. Philosophy's object of investigation is not an ideal or idealized reality in this sense. Here it may be contrasted with natural science in that physics and chemistry, for example, do sometimes abstract away features of reality – as when leaving friction out of account or assuming a perfect vacuum – whereby what is said strictly speaking holds only for reality in an ideal(ized) sense, not actual reality. By contrast, in philosophy, the *means of (re)presenting* reality, that is means of (re)presenting language use or its specific aspects may be idealized, as when describing language as used according to fixed rules or in terms of a logical calculus without claiming that language

actually is a calculus or used according to fixed rules. Wittgenstein writes:

> ... in philosophy we often *compare* the use of words with games and calculi which have fixed rules, but cannot say that someone who is using language *must* be playing such a game.—But if you say that our languages only *approximate* to such calculi you are standing on the very brink of a misunderstanding. For then it may look as if what we were talking about were an *ideal* language. As if our logic were, so to speak, logic for a vacuum.—Whereas logic does not treat of language—or of thought—in the sense in which natural science treats of a natural phenomenon, and the most that can be said is that we *construct* ideal languages. (PI, §81, emphasis original)

An earlier version of the remark expresses the same thus: 'One could at most say: "We *construct* ideal languages", in contrast to, say, ordinary language; but not, we *say* something that would only hold of an ideal language' (MS 140, p. 33/PG, p. 77). Thus, idealized philosophical concepts and concept-systems that function according to precise, definite rules may be constructed for clarificatory purposes, while acknowledging that language is actually not used thus (see also, BBB, pp. 25–6; Z, §467). What this amounts to is an explanation of a methodological idea of simplification, or abstraction or idealization, in philosophy, contrasted with idealization in science. I will return to this contrast and the notion of idealization in Section 4.

To illustrate, arguably, for example, Wittgenstein's account of meaning as use and as constituted by grammatical rules is such an abstracted or idealized model, intended only to capture specific aspects of the more complicated actual concept of linguistic meaning (PI, §43; AWL, p. 48; for discussion and justification of this interpretation, see Kuusela, 2008, ch. 4). Importantly, that this account only aims to capture particular aspects or facets of the use of 'meaning' means that it leaves open the possibility of employing other definitions or characterizations to capture other aspects of the word's use. Thus, the conception of meaning as use does not exclude the possibility of employing other different accounts of meaning in the study of the phenomena of linguistic meaning, if they suit the purposes of such a study. This is a very important consequence of Wittgenstein's conception that grammatical models constitute modes of (re)presenting the function of linguistic expressions rather than true/false claims, and are to be used as objects of comparison. On the one hand, because language is only compared with such models, the extent

to which actual language use fits them is left open. On the other hand, because to articulate a mode of (re)presentation is not to make a true/false claim about language, such models are non-exclusive. One model does not exclude others in the way in which incompatible true/false claims exclude one another (Cf. the notion of multidimensionality introduced below).

To give another example, similarly Wittgenstein's account of first-person expressions of pain as an extension of pre-linguistic pain behaviour is to be understood in this non-exclusive way. According to this account, first-person linguistic expressions replace pre-linguistic expressions, such as cries and moans, and function as manifestations of pain or avowals, rather than descriptions of inner states. Nevertheless, Wittgenstein explicitly avoids claiming that first-person pain expressions *always* function this way. There are various kinds of first-person descriptions of sensations, too – ranging from so-called secondary uses reminiscent of metaphors to more standard descriptions that we would give to a doctor. Thus, his account of first-person use as an extension of primitive pain behaviour is meant to only capture an aspect or facet of the more complex use of 'pain' (PI, §244; RPP I, §§125–6, 470).

These examples illustrate how Wittgenstein's method makes possible simplification in the sense of abstraction or idealization, whilst excluding in principle simplification in the sense of simplistic claims or theses about the objects of investigation. Moreover, Wittgenstein's account of the use of 'pain' also exemplifies the multiplicity of the methods or perspectives of grammatical description. For notably (to return to the reservation in endnote 4), here the use of the expression is not described by means of a rule. Rather, what Wittgenstein offers in *Investigations* §244 is a natural historical picture: a particular account of how children acquire pain language. But given the absence in Wittgenstein's writings of any attempts to empirically justify this account by reference to facts about language learning, evidently the account is not intended as an empirical one. Rather, just as grammatical rules can be used as models for language use, similarly a natural historical picture can be used as a model for the functioning of language in the capacity of an object of comparison. Here its role, then, is not that of an empirical natural historical account, and it differs from empirical claims, both with respect to how it is justified and generality. Firstly, the model is justified insofar as it can clarify the function of relevant expressions and resolve philosophical problems relating to them. Secondly, taken as an object of comparison, Wittgenstein's account of the use of 'pain' can

be extended further to cover sensation-language more generally, too, insofar as it helpfully clarifies the use of relevant expressions. While such an extension would be illegitimate in the case of an empirical claim without further empirical evidence, such an extension is entirely possible from the point of view of Wittgenstein's method. The extension is justified in the same way as the model was in the first place: on the basis of its clarificatory and problem-solving capacity.[5]

Finally, models that capture different aspects of the use of a word can also be used in a complementary manner to build up what may be called 'multidimensional descriptions'. (This is not Wittgenstein's term; see Kuusela, 2008, ch. 6.6 for discussion.) For example, although Wittgenstein's account of meaning in terms of rule-governed use excludes the possibility that the sound of words could have any significance for meaning (in the sense in which it can be relevant for meaning in poetry and is a condition of possibility for onomatopoeic words), this does not mean that this account of meaning could not be complemented with others that do recognize the significance of sound for meaning. However, when understood as spelling out modes of (re)presentation, rather than as claims, such accounts do not contradict but complement the account of meaning in terms of rule-governed use. (See PI, §§527–9) Similarly, descriptions in terms of rules might be complemented with natural historical pictures to make understandable features of use that a rule alone can only record but not explain, such as the grade of certainty of a judgment type or vagueness. (See Z, §374, and Wittgenstein's discussion of the notion of imponderable evidence in PI II, p. 228.) Again, this possibility of the simultaneous use of multiple models is a consequence of Wittgenstein's conception of the status or role of clarificatory models. Put forward as modes of (re) presenting the objects of investigation, models or accounts that would exclude and contradict one another when presented as true/false theses about the objects of investigation, can be used complementarily. Yet another example is Wittgenstein's characterization of language as an instrument for a predetermined purpose, and its grammar as arbitrary, not determined by such purposes. Presented as theses about the nature of language, these two statements would be contradictory (PI, §492; for discussion, see Kuusela, 2008, ch. 4.3).

Next, I will briefly discuss Wittgenstein's use of his notion of a grammatical statement to explain the role of scientific principles. This clarifies how science from his point of view requires no metaphysical grounding, and further prepares us for discussion of concept formation in psychology in Section 4.

2.3 Wittgenstein's account of scientific principles

Wittgenstein's notion of a grammatical statement can be used to characterize the role or status of scientific principles and concepts, as opposed to empirical or factual statements. On Wittgenstein's account, such concepts or principles – or grammatical statements – give expression to, or fix, norms of expression, (re)presentation or explanation (Which characterization fits best depends on the case). As he explains in his lectures, we are expressing adherence to such norms of expression, (re)presentation or explanation, for instance: (1) When a planet revolving around a star is behaving eccentrically, and we say another planet *must* be attracting it. (2) When we count two plus two apples, but find only three apples, and say, one *must* have vanished. (3) When we say that a die *must* fall on one of its six sides (AWL, pp. 15–16). He goes on to comment on the last example and then on Hertz's mechanics:

> When the possibility of a die's falling on edge is excluded, and not because it is a matter of experience that it falls only on its sides, we have a statement which no experience will refute – a statement of grammar. Whenever we say that something *must* be the case we are using a norm of expression. Hertz said that wherever something did not obey his laws there must be invisible masses to account for it. ... Hypotheses such as 'invisible masses', 'unconscious mental events' are norms of expression. They enter into language to enable us to say there *must* be causes. (AWL, p. 16, emphasis original)

Norms of expression (and so on) in the above sense might be characterized as organizational principles which enable us to see the phenomena of experience as something orderly, as exemplified by the idea that events stand in causal relations or that the universe obeys specific natural laws or functions according to the principles of mechanics. Consequently, rather than chaotic or random, reality or its specific parts appear in principle comprehensible in the sense that particular phenomena are now explainable on the basis of relevant principles, or at least we have an idea of how look for such explanations on this basis. Indeed, how deeply ingrained the norm of explanation is in us that an event *must* have a cause is illustrated by how we maintain, when searching for the cause of an event but not finding one, that we have *not found it yet*, rather than suspecting the event might not have a cause (See AWL, p. 16). Such principles may also provide us with specific types of explanation that enable us to subsume apparently exceptional cases under a more general

explanatory framework, as in the case of Hertzian invisible masses. Similarly, if in linguistics we subscribe to the methodological idea that the behaviour of a linguistic expression must always be accounted for by reference to the rules of syntax (or, alternatively, in certain kinds of cases with reference to semantic or pragmatic principles), we are adhering to, Wittgenstein would say, particular norms of expression, (re)presentation or explanation.

Notably, the characterization of relevant kinds of principles as *norms* of expression or explanation means that, according to Wittgenstein, such principles (for example, Hertz's hypothesis of invisible masses) are themselves not true or false (right or wrong). Rather, Wittgenstein maintains, they may be practical or impractical (AWL, p. 16). Accordingly, he rejects the conception of such principles as *a priori*, that is as having the status of non-empirical true/false knowledge claims.[6] He comments: 'We believe we are dealing with a natural law *a priori*, whereas we are dealing with a norm of expression that we ourselves have fixed' (AWL, p. 16). Nevertheless, this does not mean that, for Wittgenstein, it is simply a matter of stipulation and convention how we fix our norms of expression or explanation. A grammatical rule may be grounded on, and reflect, empirical regularities, as in the case of the mutual exclusion of red and green, which is not simply a matter of convention, and in such cases grammatical rules may be characterized as '... akin both to what is arbitrary and to what is non-arbitrary' (Z, §358). Similarly, as Wittgenstein explains, with the definition of the weight of iron as his example: 'It is quite possible for a proposition of experience to become a rule of grammar' (AWL, 160; cf. OC, §§96–9). Thus, what role a sentence plays is nothing permanently fixed but can change; what we first accepted as a factual empirical truth may later on serve us as a grammatical rule, for example, the definition of a natural kind. What is important for avoiding confusions, however, is to keep track of the capacity in which we are using a sentence on particular occasions. Switching unnoticed between different uses is a source of conceptual confusion.

One might describe Wittgenstein's outlook thusly: When rejecting the conception of relevant principles as *a priori*, Wittgenstein more generally abandons the explanation of the status of such principles in terms of their *origin* – as when positing empirical reality, a faculty such as Kantian pure understanding or conventions as the source of necessity or principles governing cognitive experience. For, although such principles, on Wittgenstein's account, do find their expression in conventions, that is norms we lay down for representation and explanation, this does not mean that we are at liberty to fix such principles at will. Wittgenstein

is not a conventionalist in this sense. Rather, he only emphasizes that, when elevated to the status of a grammatical rule, a statement no longer functions (in this specific context) as a factual true/false statement, but now plays a very different role. Such a statement is, so to speak, absorbed into the framework of explanation or into relevant symbolism, and becomes part of its constitution. In this sense, what is regarded as necessary, according to Wittgenstein, finds its expression in the concepts and principles constitutive of a system, rather than being expressed by means of the statements of the system.[7] Nevertheless, in the case of scientific principles, we might still say that, although the principles are not true/false as such, they can be assessed for correctness indirectly on the basis of further true/false statements that they make possible. For instance, we may ask whether they lead to correct predictions and thus enable us, further down the line, to make true statements. For, of course, it is not the case that different systems of laws concerning whatever we are trying to explain have the same explanatory power, even if it is possible to develop different such systems.[8]

2.4 Concept formation in psychology: the possibility of abstraction/idealization

Equipped with this Wittgensteinian account of scientific principles and concepts, let us turn to psychology. Clearly, the science of psychology, from Wittgenstein's point of view, requires no philosophical/metaphysical grounding on *a priori* theses about the essence or nature of the phenomena it investigates. Nevertheless, one might still wonder whether it is part of his outlook that psychology requires an analogous kind of grounding in everyday psychological concepts. This question may arise in the following way.

As noted, one way in which Wittgensteinian conceptual investigation can claim relevance for psychology is through its clarification of the use of particular everyday psychological concepts. Whenever the starting point of psychological investigation is phenomena familiar to us from everyday psychological discourse, clarity about the employment of relevant everyday concepts is important in order for us to avoid accepting readily available, but philosophically problematic (simplistic and otherwise misleading), construals of relevant concepts. Here conceptual investigation can be an indispensable help. But this still leaves open the import and authority of its clarifications. Does the possibility of clarification mean that everyday concepts – or their descriptions by Wittgensteinian philosophers! – determine how we must speak about psychological

phenomena in scientific psychology? Is it part of Wittgenstein's outlook that scientific psychology should not modify everyday concepts, if it wishes to claim that its investigations concern the same phenomena as we talk about in everyday psychological discourse? Are the concepts of psychology expected somehow to correspond to everyday concepts? No.[9] In light of the preceding account of Wittgenstein's method, it is perfectly legitimate for psychology to idealize or abstract in the sense that a concept of scientific psychology might, for example, cover only part of the phenomena subsumed under an everyday psychological concept. (Shortly I will say more about the notions of abstraction and idealization.) A scientific concept therefore might well be designed so as to capture only specific aspects of an everyday concept, similarly to how Wittgensteinian clarificatory concepts may capture only some such aspects – while still also leaving open the possibility that different perspectives expressed by different concepts might be combined into a more comprehensive multidimensional account. (I will return to this below.) For instance, a psychological study of understanding language might focus on only one of the different types of case that, according to Wittgenstein, fall under the concept of understanding a sentence and, in part, make up this concept, that is understanding a sentence in the sense of understanding what synonymous sentence could replace it, and understanding a sentence as something unique whose words are not replaceable by synonyms, as in a poem (see PI, §§531–2). Such a study need not aim to cover the different aspects of this complex concept all at once. The point can be elucidated as follows.

As Stokhof and Lambalgen emphasize in their recent discussion of the notions of abstraction and idealization, abstraction in natural science (or what Wittgenstein calls 'idealization' in *Investigations* §81 above) serves a methodological, not ontological, purpose (Stokhof and Lambalgen, 2011, pp. 11–12).[10] For example, when a physicist abstracts from friction, the abstraction has no ontological consequences in the sense that here a feature of reality is omitted from the physicist's account (as in the case of Galilean-Newtonian free fall) with full consciousness that this constitutes a simplification, and that reality does, in fact, include that feature. Because the existence of the feature in reality is not denied, it remains, in principle, possible for the physicist to take this feature into account at a later stage, that is to include it in an explanation or calculation. Here, simplification through abstraction or idealization serves specific purposes, among them the articulation of scientific laws,[11] and outlining explanations of particular actual phenomena at a stage of research where a realistic, non-idealized account is beyond

the reach of science. In such cases, we may say with Wittgenstein that the physicist is talking about reality in an idealized sense, and that the physical account only approximates how things actually are in reality. Nevertheless, because idealization as approximation of reality involves an explicit recognition that reality is not as represented, the methodological and non-ontological character of the idealization is clear on his account, too.

Similarly, abstraction or idealization in Wittgensteinian conceptual investigation is strictly methodological, not ontological. As explained above, although idealized modes of (re)presenting actual uses of language may be used in conceptual investigation (for example, calculi with definite and fixed rules), this is not to maintain that language itself is ideal in a corresponding sense. Or, as Wittgenstein puts the point in an earlier draft of *Investigations* §81, '… this 'ideal' interests us only an instrument of approximate description of reality' (MS 112, p. 94v/TS 211, p. 490/TS 212, p. 727/TS 213, p. 253v). Similar as this may sound to idealization in natural science, idealization/abstraction in conceptual investigation serves a different purpose. Crucially, Wittgenstein's aspiration is *not* to ultimately advance beyond his accounts in idealized terms and offer a non-idealized account. For him, an account in idealized terms does not merely constitute an approximate clarification provided in the absence of a proper non-idealized one. Rather, clarification by means of an ideal(ized) language is a particular method by means of which we may seek to reach 'complete clarity' about philosophical problems (see PI, §133). Importantly, it is possible to idealize or abstract in conceptual investigation, because here only those aspects of language use need to be taken into account that are relevant to the resolution of the particular problems at hand. As Wittgenstein says: '… we describe [the role of words in language] only as far as is necessary for dissolving philosophical problems' (MS 121, p. 59v; cf. PI, §182).[12] Hence, rather than merely making a problem more manageable, in conceptual investigation simplification through abstraction or idealization serves clarity in a more direct way.

The non-ontological nature of Wittgensteinian abstraction/idealization becomes especially clear when contrasted with what he calls 'sublimation' (PI, §§38, 89, 94). In this case, an ontological claim is made (similarly to idealization in Stokhof and Lambalgen, 2011, pp. 12–14). Sublimation is exemplified by the *Tractatus'* account of language, that is its conception of language as a logico-syntactical structure of precise rules that underlies actual languages and is common to them all, determining their possible uses. Here, language as the philosopher/logician's

object of investigation is understood as something ideal and abstract which we have no direct experience of, but theorize about as something that must be assumed to be there to explain the phenomena of experience, that is actually existing languages, such as English or German. Thus, the real object of investigation is conceived as something behind or beyond the phenomena of experience which themselves fail to meet the ideal in various ways. Wittgenstein comments on this: 'In reflecting on language and meaning we can easily get into a position where we think that in philosophy we are not talking of words and sentences in a quite common-or-garden sense, but in a sublimated and abstract sense. – As if a particular proposition wasn't really the thing that some person utters, but an ideal entity (the "class of all synonymous sentences" or the like)' (MS 114, p. 109/TS 213, p. 71v/PG, p. 121). Tractarian propositions are just such classes of synonymous sentences. On its account, '... what is essential in a proposition is what all propositions that can express the same sense have in common' (TLP, 3.341). Thus, what a proposition *really* is, is an abstract entity, according to the *Tractatus'* account.

However, this approach leads to a problematic separation between the concrete everyday phenomena of language and language as the postulated sublime abstract being. Now the investigation only speaks of the actual phenomena of language indirectly by speaking about that sublime entity understood as the real object of investigation. Wittgenstein explains this as leading to a dilemma of unjust dogmatism or emptiness or vacuity (PI, §131). On the one hand, in trying to make all phenomena fit an account in terms of an underlying common essence, we are in danger of doing injustice to the richness of phenomena by excluding from the class of relevant phenomena instances that should be acknowledged as part of it. In other words, we falsely exclude what ought to be recognized as linguistic phenomena from language because they do not fit our theory. On the other hand, we might respond to this problem of injustice by making the theory more accommodating and loosening the criteria for the inclusion of phenomena in the relevant class. But now we risk vacuity: nothing might be unjustly excluded, but at the cost of the capacity of the theory to distinguish instances of relevant phenomena from irrelevant ones.

Wittgenstein also compares the *Tractatus'* sublimated conception of the essence of a proposition with Freud's (early) dream-theory, according to which every dream is a wish-fulfilment dream (MS 157a, p. 56v; TS 220, p. 75/Z, §444). This is interesting as an illustration of how the problem of sublimation and problems similar to Wittgenstein's dilemma may

arise in connection with accounts intended as scientific. For instance, in Freud's theory, the recognition of dreams as wish-fulfilment dreams is expected to require their interpretation or analysis, and the theory is not applied to dreams directly or at face value. Thus, a certain amount of tinkering is accepted as necessary in order to fit the phenomena with the theory. But now, the question arises whether there ultimately is any way in which the phenomena could fail to fit the theory. If analysis/interpretation can always reveal a dream to be a wish-fulfilment dream, then the theory is able to accommodate all cases, but, the suspicion arises, only because it has now been made vacuous enough to do so. Should the theory not be able to accommodate all cases, however, the option is available here, too, to declare non-fitting cases as irrelevant, excluding them from the class of phenomena to be explained.

Sublimation, therefore, seems problematic as an approach – although, unfortunately, it may be difficult to tell it apart from idealization/abstraction in its unproblematic forms.[13] Part of the difficulty is that sublimation seems to offer a way to achieve generality in a sense desirable in science, where seemingly distinct phenomena are explained reductively as manifestations of more fundamental underlying regularities. For instance, this kind of generality is what the Freudian dream-theory would appear at first sight to achieve, until its vacuity becomes apparent, that is that there is a procedure that can always accommodate the phenomena with the theory. However, in cases of sublimation, the generality of the account is ultimately only an artefact of the mode of (re)presentation. It is achieved by misleadingly unifying the phenomena through procedures such as analysis or interpretation of dreams or other bridging principles that mediate the application of the theory on concrete phenomena. But this is to falsify the phenomena and the unity they constitute, not to achieve scientific generality (Wittgenstein's solution to this problem in the case of conceptual investigation is outlined in Section 2).

Thus, we have available two notions of non-ontological, methodological abstraction/idealization: (1) One used in natural science where features of reality may be omitted in order to spell out scientific theories or to make possible their application. (2) Another one employed by Wittgenstein in conceptual investigation, whereby aspects of complicated uses of language are captured by means of simplified models, but without aspiring to arrive at a unified overarching account in the manner of scientific theories. However, considering the possible applicability of the latter method in psychology, it might be also characterized more abstractly as a method for isolating and capturing aspects of complex

phenomena of human life, especially where the phenomena are originally spoken of in terms of everyday concepts. For, as Wittgenstein's discussions of everyday psychological concepts may be taken to show, the expectation of their neat definability is not well founded. We cannot *assume* the concepts constitute simple unities that can be captured by means of overarching definitions.

Which notion of idealization/abstraction would best suit the needs of psychology? It might employ both approaches. However, given the complexity of everyday psychological concepts and phenomena which we speak about in their terms, the method of Wittgensteinian abstraction/idealization might turn out useful for psychology in the following sense.[14] Whenever it is not possible or desirable to explain away the complexity of a class of psychological phenomena (such as remembering or thinking), or when such attempts lead to problematic simplification in the sense of sublimation, Wittgensteinian abstraction/idealization may offer a way to make the phenomenon, or aspects thereof, accessible to research. Rather than seeking to explain a class of relevant psychological phenomena with reference to their presumed underlying unifying core – whereby the problem of sublimation is imminent – psychology might usefully approach its objects of investigation in a more piecemeal Wittgensteinian fashion, treating separately different aspects of those objects. This approach might then not only make the task more manageable through division into sub-problems and sub-areas of research, but the approach could have fruitful consequences also in the following sense.

Recall the notion of multidimensional Wittgensteinian descriptions, introduced in Section 2. As explained, this notion refers to the possibility of employing different (and different types of) Wittgensteinian abstracted/idealized descriptions of language use simultaneously to capture different aspects of complex concepts. Similarly, however, such multidimensional accounts might be developed in psychology, whereby an account that focuses on one aspect of a complex psychological phenomenon could be complemented by other accounts that aim to explain some different aspect of the phenomenon or seek to explain it from a different perspective, that is in the framework of a different set of concepts. Importantly, it would seem possible to achieve through multidimensionality comprehensiveness of accounts of psychological phenomena without assuming their unity in the sense that all their aspects could be captured in terms of one unifying account, perhaps merely an illusory sublimated underlying explanatory entity.

More broadly, it seems, in principle, possible that different systems of scientific concepts and principles[15] could complement one another in this manner, without assuming that they all must be part of a single general consistent system, or perhaps reducible to a single conceptual and explanatory framework considered more fundamental than others. The latter way of thinking is exemplified by the aspiration to reduce discourse in psychological terms to discourse in neurophysiological terms. But rather than assuming that the relation between concept-systems such as the psychological and neurophysiological ones must be one of reduction or systematic correlation, it seems possible not to assume this, whilst still maintaining that the two discourses are both concerned with the same phenomena. For it seems we can understand the two discourses as talking about the same phenomena from different angles already because we know, in some basic enough sense, that there is a connection between mentality and having a brain. That connection, however, need not be anything systematic in the sense that it should be possible to spell it out in terms of any rules of correlation between the vocabularies of the two discourses (cf. Z, §§610–3). This, at any rate, seems to be Wittgenstein's view: We are assuming a too-primitive conception of the function of everyday psychological locutions, if we assume them to correspond neatly to neurophysiological locutions (see introduction and RPP I, §906; Z §611). To question the expectation of such a simple connection between the two discourses or conceptual frameworks is not to say that there is no connection at all, but to open up a space for the exploration of more complex ways to understand their relation. This can be understood as exemplifying the idea of what I have called the 'multidimensionality' of accounts.

Finally, the preceding account of Wittgensteinian abstraction/idealization also provides us with an answer to the question raised earlier; whether psychology needs to stick to everyday concepts, insofar as it wants to maintain that its investigations concern the same phenomena as we speak about in everyday psychological discourse. The answer is negative. The concepts of psychology might idealize/abstract analogously to Wittgensteinian clarificatory concepts, while retaining a clear connection with the phenomena spoken of in everyday psychological language. From the perspective of Wittgenstein's method of conceptual investigation, to deviate from, or agree with, everyday concepts is not an all-or-nothing matter. It is not the case that anything short of perfect correspondence with everyday concepts would mean speaking about something else rather than what we speak about in everyday life.[16]

Notes

1. I have replaced Wittgenstein's comma with a semicolon. References to his *Nachlass* are by manuscript or typescript number. Unless a reference is given to an extant translation, the translation is mine. Abbreviations used for Wittgenstein's works are given in the bibliography.
2. See RPP I, §950 for a brief discussion, including how conceptual investigation can give a new direction to science. The notions of mode of expression and (re)presentation are to be understood broadly. According to Wittgenstein, relevant kinds of confusions do not arise merely in the medium of word-languages, but may also arise in connection with pictorial or mathematical modes of (re)presentation. See PG, p. 194 for example, of an impossible design for an engine that Wittgenstein uses to illustrate the notion of a conceptual confusion, and pp. 374–5 for an example connected with mathematics.
3. No doubt, philosophers to whom psychologists turn when looking for ways to conceptualize their objects of investigation are partly to blame for the adoption of misleading concept-construals in psychology. One example is the idea that the function of verbs such as 'believe' or 'hope' or generally verbs regarding so-called propositional attitudes, is the description or ascription of mental states (that are then the basis of prediction of behaviour, for example). (See Fodor, 1987.) Arguably, although I cannot discuss this issue here, this is far too simple an account of the function of relevant verbs, and not the basis of folk-psychology of any folk outside philosophy (and perhaps psychology) departments. Accordingly, when eliminative materialists reject folk-psychology thus construed, they are, arguably, not rejecting everyday psychological understanding, but merely a philosophical mis-description the latter (see Churchland, 1989).
4. This characterization involves a simplification to which I will return. For the distinction between rules and factual statements, the difference of conceptual investigation from metaphysics and the danger of the relapse of grammatical investigation into metaphysics, see Kuusela, 2008, ch. 3.
5. Descriptions in terms of natural historical pictures and grammatical rules do not exhaust Wittgenstein's methods of clarification. For example, a method that partly overlaps with these two, but also has distinct features, is the method of simple language-games. (For discussion, see Kuusela, forthcoming.) For a list of Wittgenstein's methods, see Kuusela, 2008, p. 270.
6. This is part of Wittgenstein's rejection of the *a priori-a posteriori* distinction as a distinction between different types of *knowledge* (see TS 220, §91/Z, §442). The distinction thus construed, according to him, makes the difference between empirical or factual and grammatical statements too small. Grammatical statements are not a special kind of knowledge claims, but have a different role.
7. In this sense, a psychological model of mind or thinking, according to Wittgenstein, would be part of the symbolism of the theory, like a mechanical model constitutes a part of a theory of electricity. (See BBB, p. 6; PG, p. 106)
8. For Wittgenstein's conception of necessity as expressed by grammatical statements, and how his view leads beyond the opposition between realism and idealism (including linguistic idealism or constructivism), see Kuusela, 2008, ch. 5.

9. There are various issues relating to this question that cannot be discussed here. For a critique of the interpretation that philosophy, according to Wittgenstein, could (or even aims to) tell us how we must think about matters, and an alternative interpretation of Wittgenstein's view of the importance of seeking agreement in philosophy, see Kuusela 2008, chs 6.3–6.4. For a reading of what Wittgenstein means by leading words back to everyday language that does not assume everyday use to constitute a standard for sensible language use, see ch. 7.3. By contrast, for a Wittgenstein-inspired discussion of experimental work on neonatal imitation that seeks to criticize psychologists for departing from the everyday use of the word 'imitation' on the grounds that this leads to 'generating nonsense', see Tissaw (2007). Here nonsense is characterized as a case where '... words are used not in accord with uses to which we are accustomed' (Tissaw, 2007, p. 222). But whilst any half-baked departures from everyday uses that let everyday senses enter inferences unnoticed are certainly to be criticized, just as any inferences involving equivocation, Wittgenstein is explicit that departure from everyday use or an established use as such is not philosophically problematic (See RPP I, §§548–50).
10. The terms 'abstraction' and 'idealization' do not have established uses in connection with (the philosophy of) science. For example, Stokhof and Lambalgen (2011) and Jones (2005) use them in roughly opposite meanings.
11. This is what Galileo did when spelling out the law of inertia, according to Nowak (2000), and Maxwell when employing idealized models of electro-magnetic phenomena to develop the mathematics for it, according to Nersessian (2005). Interestingly, Wittgenstein compares his own method with Boltzmann's similar employment of physical analogies/models to develop mathematics for the behaviour of gases. Discussion of this issue is beyond the scope of this chapter, however. (VW, pp. 288; cf. MS 111, p. 120;TS 211, p. 73)
12. For fuller quotation and discussion, see Kuusela, 2008, pp. 79ff.
13. Stokhof and Lambalgen (2011) raise the question of whether the account of language in generative linguistics constitutes an idealization in a problematic sense corresponding to sublimation, and opposed to abstraction in the sense of omission of features.
14. This method might be useful in social sciences and humanistic disciplines more widely, but the discussion of this is beyond the scope of this chapter.
15. Here I am only talking about systems of scientific concepts and principles that are successful by relevant criteria, whatever they are, not just any systems that might be proposed. I am also assuming that the status of such principles and concepts is understood in the Wittgensteinian manner outlined in Sections 2.2 and 2.3.
16. I am grateful to Rupert Read and Angus Ross for comments, as well as to the editors of the present volume, Tim Racine and Kate Slaney.

References

P. M. Churchland (1989) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (Cambridge, Mass.: The MIT Press).

J. Fodor (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, Mass.: The MIT Press).

M. R. Jones (2005) 'Idealization and Abstraction: A Framework' in M. Jones and N. Cartwright (eds) *Idealization XII: Correcting the Model: Idealization and Abstraction in the Sciences* (Amsterdam: Rodopi).

O. Kuusela (2008) *The Struggle Against Dogmatism: Wittgenstein and the Concept of Philosophy* (Cambridge, Mass.: Harvard University Press).

—— (forthcoming) 'The Method of Language-games as a Method of Logic' in *Philosophical Topics*.

N. J. Nersessian (2005) 'Abstraction via Generic Modelling in Concept Formation in Science' in M. Jones and N. Cartwright (eds) *Idealization XII: Correcting the Model: Idealization and Abstraction in the Sciences* (Amsterdam: Rodopi).

L. Nowak (2000) 'Galileo-Newton's Model of Free Fall' in I. Nowakowa and L. Nowak (eds) *Idealizations X: The Richness of Idealization* (Amsterdam: Rodopi).

M. Stokhof and M. van Lambalgen (2011) 'Abstractions and Idealizations: The Construction of Modern Linguistics', *Theoretical Linguistics*, 37(1/2), 1–26.

M. A. Tissaw (2007) 'Making Sense of Neonatal Imitation', *Theory and Psychology*, 17(2), 217–42.

3
Pictures of the Soul

Joachim Schulte

> Religion teaches that the soul can exist when the body has disintegrated. Now do I understand what it teaches? – Of course I understand it – I can imagine various things in connection with it.
>
> After all, pictures of these things have even been painted. And why should such a picture be only an imperfect rendering of the idea expressed? Why should it not do the same service as the spoken doctrine? And it is the service that counts. (PPF, §23)

Most of the remarks Wittgenstein wrote between the completion of the typescript of his *Philosophical Investigations* (1945–1946) and his visit to America in the summer of 1949 are connected with his investigations into our psychological concepts. Obviously, the words we use in using these concepts do not form a well-circumscribed class of linguistic expressions. Nor is Wittgenstein interested in supplying criteria that might help us decide whether or not a given expression is part of our psychological terminology. He does not examine the boundaries between different disciplinary vocabularies. What he does deal with, however, are certain complicated relations between our psychological concepts and facts of nature: 'Indeed the correspondence between our grammar and general (seldom mentioned) facts of nature does concern us' (RPP I, §46; cf. PPF, §§365–7). But this kind of concern is not of a scientific nature: it is not directed at possible causes of our having these (rather than other) concepts, even though comparing our actual concepts with possible concepts of a different kind forms an essential part of Wittgenstein's philosophical enterprise. The following considerations are meant to throw some light on a particular aspect of Wittgenstein's idea of a conceptual investigation. At the same time, they are meant

to illustrate two points: first, for Wittgenstein there is no such thing as a purely conceptual (or 'grammatical') investigation; second, there are certain questions that arise again and again in his early as well as his latest writings, and questions of picturing and representation are chief among these.

3.1

It is often claimed that the following two brief remarks from section iv of what used to be called 'Part II' of the *Investigations* epitomize Wittgenstein's view of the relation between body and soul: 'My attitude towards him is an attitude towards a soul. I am not of the opinion that he has a soul' (PPF, §22), and 'The human body is the best picture of the human soul' (PPF, §25). What cannot be denied is that these two remarks deserve our attention. At the same time, they stand in sore need of interpretation on account of the fact that they may suggest various misunderstandings which are not easy to correct.

One reason why these remarks are often misunderstood lies in the history of 'Part II'. It tends to be treated as if its status were similar to that of the former 'Part I', which, however, is the only part of the published book which deserves the title *Philosophical Investigations*. The 693 remarks collected under this title form a text that was re-assembled and revised several times by its author. Even though it no doubt falls short of what he intended to achieve, it must have appeared sufficiently organized to him to call it 'my book' (OC, §290). Nothing of this kind can be said of PPF. The status of this text – based on a hypothetical typescript derived from a selection of remarks taken from two typescripts and a few manuscripts written between 1946 and 1949 – is questionable for the reason that the only extant testimony is a number of handwritten pages collected in a loose-leaf folder.[1] Of some of these pages, we can say with certainty that their present order is not the original one. It is also obvious that this material is an early, unfinished and, at most, partially ordered selection of remarks. Nothing is known about a plan the author may have followed in selecting these notes, so there is next to no textual basis one might appeal to in trying to reconstruct what Wittgenstein was up to when he worked on this compilation.

This should suffice as a general sketch of the problems surrounding this fragmentary text. As regards section iv (as it was styled by the original editors), it remains to be pointed out that its eight remarks (now numbered §§19–26 of PPF) were written down by Wittgenstein on the recto and verso pages of one and the same sheet of paper, whose second

page was not completely filled. It may well be that the typescript (bearing the number 234 in von Wright's catalogue of Wittgenstein's papers) used as copy for the typesetters would have helped us to gain a better understanding of Wittgenstein's intentions. But the typescript was lost, and the extant material cannot, for example, tell us whether section iv was meant to continue, or to be continued by, other remarks. At any rate, the remarks forming section iv were selected by Wittgenstein from two different manuscripts, the earlier of which was written more than two years before the later passages. Sections 23–6 of PPF were written in Cambridge on 19, 20 and 31 August 1946, while §§19–22 were put down in Dublin on 27 November 1948 (cf. LW I, §§318, 322–4). What should also be taken into account is the fact that the four later remarks were written only a few months before the selection was made – in contrast to §§23–6, which were taken from much older material. In this case, the selection was made, not from a manuscript, but from a typescript dictated in autumn 1947 (see RPP I, §§265, 279, 281, 345).

3.2

Now, some readers may wonder what all this information is meant to be good for. After all, the sense of our two chief remarks (§§22, 25) appears to be pretty clear and not to require additional explanations. This is a view that seems to be fairly widespread: It happens again and again that these observations are quoted as if they were absolutely free-standing and permitted exactly one possible reading, or as if they were conclusions drawn from an entirely self-evident process of reasoning. I myself, however, feel that the sense of these two remarks (not to mention the other observations of this section) is far from obvious. Here it may be useful to look at the original manuscript context and try to find what sense can be made of our remarks if seen within this framework. If an interpretation given on the basis of an examination of this material agrees with what the reader believes to be the correct reading anyway, the result will speak in favour of his intuitive powers.

What does it mean to say that a, or the, 'human body is the best picture of the, or a, human soul'? Two problems stand out and need to be answered before getting down to giving one's interpretation the finishing touches: (1) If one takes the actual words of this passage seriously, there can be no doubt that body and soul are sharply separated. This, however, is difficult to harmonize with all the other things we know, or believe to know, about Wittgenstein's views. (2) The suggestion appears to be that there are a number of possible pictorial representations of the, or

a, human soul, and that among these the representation relying on the body is the best choice. Is there a criterion which could be appealed to in support of this judgment? And in what sense is it supposed to be permissible to call the body a picture, in particular a picture of the soul? These questions indicate problems of interpretation which cannot really be left out of account and which stand in urgent need of an answer.

If one looks at the manuscript context of the early version of what became §25 of PPF (= MS 144,[2] p. 7v) taken down on 20 August 1946, one will immediately notice the following fact. The remarks written on this day can be divided up into two completely different groups of observations: The first 7 remarks (MS 131, pp. 72–6) concern questions of the analysis of if-then sentences considered in the light of the law of excluded middle and terminate in a reference to puzzles (attributed to Peter Geach) concerning the difficulty of giving a general account of sentences like 'Have you stopped beating your wife?' (RPP I, §274). The second division (MS 131, pp. 76–80; cf. RPP I, §§275–81) begins with remarks on various mental attitudes (desire, yearning, hope and belief), where, among other things, a certain contrast between attitudes like desire and yearning, on the one hand, and hope as well as belief, on the other, comes into play. Here, using the general term 'attitudes' is hardly more than provisional, a mere expedient, for Wittgenstein himself employs the striking phrase 'behaviour of the soul', that is, he uses the signature expression of behaviourism to characterize attitudes of the soul. In other words, in order to articulate the idea to be examined, he uses a word combination which, at first blush, seems quite outrageous: 'Desire is a kind of behaviour of the mind, the soul, towards an object.' 'Desire is a state of the soul which refers to an object.' At this point, it does not matter whether Wittgenstein would approve of these or similar sentences. They are examples of certain philosophical attempts to get a grasp of the attitudes alluded to; and the evidently contrived formulations make it clear that, on the one hand, these attempts are, to a high degree, theoretically charged and that, on the other, they are meant to capture and illuminate aspects which are deeply rooted in our thought and feeling.

Yearning, it seems, is a particularly apt example of the supposed referentiality of a mental state. There is no problem about sketching a picture of such a state of affairs: on the one hand, there is a human subject, and on the other there is the desired object in front of his eyes. This idea, which strikes us as immediately appealing in cases of direct confrontation between a human person and the object of his yearning, is then apparently easy to apply to cases where the object is not present but

represented first by a picture and then by a mere visual image of the object: 'If [the object] is not there in front of us, perhaps its picture goes proxy for it, and if there is no picture there, then a visual image' (RPP I, §275). Wittgenstein's word is *vertreten* – the crucial term of the *Tractatus* theory, according to which names are representatives of objects. The beauty and immediate persuasiveness of this idea is now brought to bear on our notions of attitudes like desire and yearning; and again, the concept of a picture does its work in virtue of the ease with which we can grasp the idea of one thing's going proxy for another. It seems that the differences between the real thing, a drawing or a photograph and a visual image hardly matter: Any old thing can stand in for another as long as we adopt the right attitude towards it. And it is this sort of attitude which is here vividly described as a kind of behaviour of the soul towards its object.

Wittgenstein's next move (in the same remark) is highly cunning. He himself uses the idea of one thing's going proxy for another and substitutes 'body' for 'soul': *au fond* it is the behaviour of a body which we have in mind when we speak of the behaviour of the soul. (We must not forget, however, that normally we do not speak of the behaviour of a soul; it was Wittgenstein who introduced this form of expression to make the philosopher's way of talking about attitudes more naturally appealing.) It becomes clear now that it is the familiar notion of the behaviour of a human body which lends the idea of the soul's behaviour whatever intelligibility it may have. Only that normally we do not think of the body as different from, or opposed to, the soul in this sort of context; we simply think of the human being as bearer of the behaviour in question while questions of soul vs. body just do not arise. So, we are brought to see that our understanding of all the critical terms under discussion is in one way or another parasitic on our understanding of the ordinary notion of a human being's behaviour (A similar view is suggested by well-known remarks like PI, §§281 or 283).

But this by no means exhausts the points Wittgenstein raises in the first remark of this sequence. As we have seen, his last move was one of actually using the idea of *Vertretung*, which previously had been alluded to in his description of how easily we can slide from real thing to picture and visual image.

Now he proceeds to use the idea of a picture in claiming that the soul's behaviour towards a visual image is precisely what could be made vivid by drawing a picture of a man's soul which leans with a longing gesture towards a painted picture of an object. But of course, there is no way of drawing a picture of a man's soul which dispenses with drawing a

picture of his body: Only by drawing a picture of a man's body can you draw a picture of his soul. And the depicted embodied soul's leaning towards a painted picture indicates how it works: Context, tradition and conventions of painting and reading paintings join forces to make us see that the soul's behaviour is not really directed towards a painted picture but towards something that is of a nature more closely related to that of the soul itself, viz. a visual image. This train of thought as presented in Wittgenstein's remark is splendidly organized. It brings out how our philosophical way of thinking tends to work in such cases, and at the same time it exposes the mechanism that lies behind it. Evidently, this way of thinking has a mythological element. But even though Wittgenstein's description helps us see its mythological nature, he does not actually condemn it; rather, he is concerned to show us how natural it is.

'In the same way', Wittgenstein observes, 'one could represent what it is like when a man does not give any facial expression to his desire, while his soul craves it'.[3] Clearly, the 'same way' needs to be spelled out. It surely would not be enough to draw a picture of an expressionless face with eyes directed at a certain object or at a picture of this object. Nor would it be sufficient to draw a picture of a person with an expressionless face in whose heart or head a picture of the craved object is to be seen. This would not help because, even if it were taken as a representation of some attitude, one would not be any the wiser about which kind of attitude was meant. So, what is required is a representation of the soul's craving – a craving which is hidden behind a stony or indifferent face. And, of course, we know how to solve this problem: we simply draw a second human figure and somehow make it clear that this is the soul.[4] By now we know that we cannot draw a soul without drawing a body, but that raises no great difficulty. There are many conventional or obvious ways of making it clear that a certain figure is meant to be the soul belonging to a given body. And this second figure, the portrayed soul, will then have to be connected with the object whose being desired by the person with the indifferent face requires for its pictorial representation this clear-cut split, whose meaning we all understand – never mind whether we are behaviourists, mentalists, agnostics or what have you.

The next two cases which Wittgenstein considers in his manuscript (but not in the typescript drawn from it)[5] are helpful, especially as they may be apt to muddy the waters. (From Wittgenstein's practice, one gains the impression that sometimes muddying the philosophical waters is just as important as attempts at clarification.) The first case is that of hope, the second case concerns fear. Hope, Wittgenstein says, is similar

to an unexpressed desire – only more difficult to represent. It is difficult to represent because it resembles belief, and hence is a 'silent possession of the soul'. It is similar because you need a means of indicating that it is, not the body, but the soul, the mind, which stands in a certain attitude towards what is believed or hoped. The attitude itself can be expressed by using conventional or obvious ways of depicting possession, for instance by showing a figure clutching the belief and pressing it against his heart. But that leaves the difficulty of representing the belief or hope, and this difficulty is most easily overcome by resorting to writing – a written statement which the figure in question 'carries in his soul and which assures us that things are thus and so'.[6]

Fear, on the other hand, is easy to represent. According to the relevant remark of Wittgenstein's, all you need to do is to depict 'the behaviour of a fearful human being confronted with the object of his fear'. That is, there is no need to introduce a second figure to represent the attitude in question, nor is there a special problem about how to indicate its object. At this point, I suppose, Wittgenstein is simplifying matters. In cases where fear shades into worry or angst, the matter will be more complicated and raise questions about how to depict non-obtaining states of affairs. And, come to think of it, their depiction may call for clarifying by pictorial means that it is the mind (and not the body) which does the entertaining of such states of affairs. So, while Wittgenstein is surely right in observing that standard cases of fear can be represented by depicting expressive behaviour of human figures and corresponding objects of fear, there will no doubt be more complicated cases whose portrayal will require us to resort to similar means as those mentioned in our descriptions of earlier examples.

3.3

As we have seen, practically all the remarks of the sequence we are dealing with concern problems of pictorial representation, and the word 'picture' (Bild) is actually used quite frequently. Another topic which pervades these remarks is the idea of the human soul as distinguished from the body, and in this context it has turned out that our philosophical ways of talking about 'psychological phenomena' (*psychologische Erscheinungen*) are, on the one hand, parasitic on our ordinary ways of talking and, on the other, dependent on certain paradigmatic pictures, or models, which, while not exactly part of our ordinary language, seem to be deeply rooted in this language. Wittgenstein uses a number of characteristic expressions in his observations on this aspect of the

matter, and some of them come to the fore in the remaining remarks of our sequence, which form the centre of our discussion.

Here, Wittgenstein mentions certain forms of expression which we tend to use in describing hope or yearning. Thus, yearning may be felt as a weight loaded onto my verbal expression by my heart. Now, however, it is not so much the image itself which Wittgenstein is interested in, but rather the fact that we want to say such things. 'Wanting to say it' is one of the characteristic expressions Wittgenstein tends to use in this context; an expression's 'forcing itself' upon the speaker is another: I may feel inclined, or tempted, to use certain words; perhaps I am unable to resist a tendency to employ a given picture; sometimes it comes naturally to me to say certain things; at other times, I cannot help using a particular figure of speech. Now one may wish to reply to this that these are merely subjective feelings. Why should my impression that I cannot help using a certain phrase be important to people who are not particularly interested in me? The fact that I cannot help it is significant, Wittgenstein suggests, because the picture I am inclined to use is likely to be a 'picture of some important kind of human behaviour'. My inclination indicates that it lies in our nature to conceive of this behaviour in a way illustrated by pictures and expressions we cannot help using in the relevant sort of situation.

Hope, for example, is said to be alive in my chest. Thought, on the other hand, is located in my head. And this, Wittgenstein insists, is no mere whim, nor is it an hypothesis. It is something we actually experience in thinking. Just look at people's behaviour: the thinker 'clutches at his head; he shuts his eyes in order to be alone with himself in his head. He tilts his head back and makes a gesture to signal that nothing should disturb the process in his head.' And to this description, Wittgenstein adds the rhetorical question of whether these are not 'important kinds of behaviour' (RPP I, §278). Obviously, they are important, if what we are concerned about is understanding human beings and their behaviour.

The picture of thought in the thinker's head makes Wittgenstein wonder, and in the light of his earlier reflections on pictures which, in certain situations, we find it natural or inevitable to invoke, he raises the following question: If the picture of a thought in one's head is so compelling, why shouldn't the picture of a thought in one's soul be even more powerful? (RPP I, §279; PPF, §24.) The answer to this question is not obvious. As a matter of fact, we do not use the picture of thought occurring in one's soul. In what way would it be a picture? After all, the soul (in the sense of 'mind') is the place where thinking takes place – if it makes any sense to talk about a location of thinking at all. And if,

in spite of these doubts, we decided to use this picture, how would we go about drawing it? Presumably, we would draw a picture of a human figure, or part of a human figure – For example its head, and proceed to inscribe the thought or its representation in the place which seems most natural: maybe the head, maybe the heart – it will depend on the sort of thought concerned.

So, Wittgenstein's question has led us round in a circle. If what you want is a picture, you will have to use a means sufficiently different from the object you wish to portray. Coloured lines and patches on a flat surface can function as parts of pictures of people, scenes and activities; so can toy cars, match boxes and matches. You can use people in the style of a *tableau vivant* to represent people; but in such a case, it may prove urgent to use explicit means of forestalling possible doubts about the pictorial character of the *tableau*. What you cannot do, however, is to produce a portrait of person X by means of exhibiting person X. (Which does not mean, of course, that person X cannot give a pictorial account of a certain scene from her or his life, nor that he or she cannot instantiate, and by this means depict, a characteristic feature of mankind or a group of human beings.) At any rate, these considerations have helped us to appreciate two near-truisms: first, there is no straightforward way of introducing the soul into the drawing of a picture; second, pictures seem to be the only helpful way of approaching the soul – of getting a vivid idea of what people mean when they speak of the soul.

3.4

This is the background of the two brief remarks concluding our sequence: 'What better picture of believing could there be than a human being who, with an expression of conviction, says, "I believe…"? A human being is the best picture of the human soul' (RPP I, §§280–1). Notice that here we are dealing with two best pictures: 'What better picture could there be…?' and '…is the best picture of…' A human being who says 'I believe…' and pronounces these words with a facial expression of confidence is the best picture of believing. It may even be good enough to play a role in the context of teaching certain uses of the word 'believe'. Of course, a painting or a photograph of a man may be equally suitable for conveying the idea of conviction, but here we encounter the problem mentioned above about representing the belief itself: We may have to resort to written or recorded words, and that may well interfere with the persuasiveness of the representation as a whole.

The second one of the two 'best pictures' mentioned is a man, a human being, who is said to be the best picture of the, or a, human soul. I think we have now gathered enough material to contextualize this remark. First, we have to be clear about why we want or need a picture of the human soul in the first place. There may be a number of reasons for requesting such a picture, but they all seem to stand in need of a good deal of elucidation. One plausible candidate would be the case alluded to in a remark written on the day before Wittgenstein concluded the sequence we have been looking at. Here he speaks of a 'soulful facial expression' and the fact that such an expression can be painted. This would be a context where you may find it worth stating that no better picture of the soul can be painted than one portraying a human being (and then you could go on to specify what kind of portrait would be particularly useful). Another context may be that of a discussion of certain religious or theological views of a soul. Here, the way in which you will be able to specify which sort of portrait suits your purposes best will depend on the ideas you wish to explain or defend. Generally speaking, no kind of picture will be better than one depicting a human being. Even if what you wish to portray is the soul of a god, you will produce a picture of a human(-like) being; and if you want to produce a picture of the soul of an animal, you will humanize its features and by these means convey the idea of its having a soul of a certain kind.

Perhaps it is worth emphasizing that Wittgenstein is not talking about a given sort of entity called our 'soul', of which he then proceeds to claim that there is no better way of picturing it than by way of producing a picture of a man or even the man himself. He assumes that we do have ideas of what a soul is or may be, and in his view many of these ideas are, as we have seen, dependent on certain ways of talking about 'psychological phenomena' and our attendant techniques of portraying these phenomena. He underlines the fact that certain descriptions and representations of such phenomena are deeply rooted, not only in our concepts, our forms of speech, but also in our nature as human beings: these are ways of describing things which come naturally to us, which we cannot help using and which, in virtue of their very instinctiveness, are apt to reveal something about our 'mindset'. And if one wishes to find out about these ways, it will be useful to examine forms of expression and kinds of picture we tend to use in situations where we feel that these forms 'force themselves upon us'.

But if you are confronted with the task of drawing a picture, of producing a vivid image, of the soul – that is, of what we or other people mean by speaking of a soul, then you will find no better means of solving

this problem than by producing a picture of a human being and pointing out which features of the pictured being contribute to expressing the soul. If, on the other hand, there is not enough shared cultural background for raising questions about what another person may mean by speaking of the soul, no picture will help us grasp what he is talking about. Useful pictures[7] may be simple enough and very vivid, but they come in useful only if there are questions, or a context for questions, that could be answered by producing such pictures. There may be tough philosophical problems that can be solved by pointing to one simple picture, but it needs a philosopher (and his background) to worry about the sort of problems answerable in this fashion. Only in the context of a philosophical question can a remark like 'A [picture of a] human being is the best picture of the human soul' help to make a point.

3.5

This interpretation of the relevant remarks may be quite speculative, and I do not expect all readers to agree with most of what I have said. But no matter how readers will respond to it, they will have to take into account that this is an interpretation which allows for the manuscript context and hence for the context of ideas and worries that prompted Wittgenstein to write these words. If, on the other hand, you look at the remark on the best picture of the human soul as it is printed in the *Investigations*, you will find it hard to make much use of the text surrounding it there. But, of course, now that we are looking at the printed page of the *Investigations*, we notice a remarkable difference between the version I have quoted and the version as given in the book. It does not say that a human being is the best picture; it says that a, or the, human body is the best picture of the human soul. This modification of the original wording[8] surely makes a difference. It introduces a sharper contrast between the two terms – 'body and soul' – and it emphasizes possible connections with religious or theological views, which are alluded to in §23 of PPF. To speak of a human being can be understood as referring to such a creature in its entirety, and it may even be understood as referring to a creature that cannot clearly or helpfully be divided into two separate substances. This is a reading that §25 of PPF seems to exclude: The sentence employs the classical – philosophical or theological – terminology standardly used to articulate a dualist point of view.

But even if you leave the original manuscript context out of account, it is possible to read our sentence in a way which removes it from the

concerns of classical-all-too-classical discussions of dualism and related conceptions. One impressive example is the interpretation developed by Stanley Cavell in Part Four of his *Claim of Reason*, from which the following quotation is taken:

> The idea of the allegory of words [that our attachment to our words is allegorical of our attachments to ourselves and to other persons] is that human expressions, the human figure, to be grasped, must be read. To know another mind is to interpret a physiognomy, and the message of this region of the *Investigations* is that this is not a matter of 'mere knowing'. I have to read the physiognomy, and see the creature according to my reading, and treat it according to my seeing. The human body is the best picture of the human soul – not, I feel like adding, primarily because it represents the soul but because it expresses it. The body is the field of expression of the soul. The body is of the soul; it is the soul's; a human soul has a human body. (Is this incomprehensible? Is it easier to comprehend the idea that it is the body which has the soul? (Cf. §283) It does seem more comprehensible (though of course no less figurative) to say that this 'having' is done by me: it is I who have both a body and a soul, or mind.) An ancient picture takes the soul to be the possession of the body, its prisoner, condemned for life. (Cavell, 1979, pp. 356–7; words in square brackets, p. 355)

Not only does Cavell succeed in connecting our passage with the highly relevant remark PI, §283; he also manages to bring in questions of 'reading' physiognomies and emphasizes expression (while minimizing representation). Both points can be accommodated, I think, within the framework of the account given above, especially if we do not forget the manifold conventional ways and means of portraying and reading expressions. What is missing from Cavell's sketch, however, is an awareness of the crucial point that the body's qualification as best picture of the soul is dependent, first, on our willingness to engage in playing the body-soul game and, second, on our going along with the practice of playing this game to an extent which allows us to raise questions in whose terms a picture of the soul can appear to promise an answer.

These insights have to be respected, and respecting them will require taking a certain distance from the game while playing it. In some parts of Wittgenstein's writings, irony helps to provide the distance needed. But there is no irony to be detected in 'this region of the *Investigations*'. A possible reason for its absence may be connected with a point much

exploited by Wittgenstein and mentioned before, viz. the fact that these ways of talking and picturing come naturally to us, and that we cannot help making use of them if we wish to engage in talking about the soul and picturing it at all. Perhaps, irony would not be an adequate method of indicating distance if the means of expression under discussion are not means of our choice.

3.6

The idea that certain forms of expression and ways of picturing force themselves on us and are so compelling that we cannot help using them plays various roles in Wittgenstein's thought. Here, I want to connect it with the only thing I wish to say about the other oft-quoted remark from our section of the *Investigations*: 'My attitude towards him is an attitude towards a soul…' There is a certain risk here of misreading the word 'attitude'. Wittgenstein's German word is 'Einstellung', and his use of this word is exposed to the same risk as the word 'attitude' used in the translation. The risk is that in speaking of attitudes we tend to think of changes and differences of attitude: His attitude used to be left-wing, but he has become a reactionary – so he changed his political attitude; he is a very devout man, but his brother is a committed freethinker – so their religious attitudes are different. The risk lying in this quite natural sort of reading of 'attitude' is noted by Cavell when he wonders whether, in making his remark about my attitude towards another person, Wittgenstein is subject to a certain prejudice: after all, 'my attitude is a state of just this organism; it is a passage of just my history; a passage I might find myself in, or take, at any time, regardless of the circumstances' (Cavell, 1979, p. 398). I think if we read Wittgenstein along these lines, we would seriously misunderstand him. The attitude he has in mind is an ingrained way of responding to the other person: I cannot help perceiving him as a soul – not as a body plus soul, but simply as a soul. That is, in my eyes he is a being regarding whom I would not dream of treating him as I might treat a robot or a cockroach. In this case, my attitude is not something I may change tomorrow without running the risk of becoming a different sort of creature. It is an expression of natural inclinations and a history which is not just mine, but the history of a culture, a large segment of the history of mankind. In other words, this attitude is part of the natural history of human beings (PI, §415). But the natural history of human beings (as distinguished from that of cats and dogs) is, at the same time, an (admittedly very partial) history of our culture: giving orders, asking questions, telling stories and having a chat (PI, §25) may belong to a

different chapter from that which deals with our attitude towards the other soul, but they are, none the less, chapters of the same book.

To the extent my attitude towards the other person is an attitude towards a soul, it can neither change nor be changed without entailing a change in my nature. There is no *perhaps* and no *as if* about it.[9] It may also be worthwhile to stress that this has nothing whatsoever to do with aspects and changing aspects. While expressing that for a human being this is an, as it were, preset attitude, Wittgenstein manages to gain a certain distance from it – the distance that is necessary to show that this is a reflective insight. And his means of gaining this distance is by using the word 'soul' in speaking of my attitude towards the other person. Another writer might wish to say that what Wittgenstein means is that my attitude towards the other is (or cannot help being) an attitude towards a human being (twice underlined). But this would not be the same thing: Wittgenstein's way of putting the matter cannot be reduced to a thesis, be it ever so emphatic. He expresses what he wants to say by using a picture that cannot be replaced by a different kind of thing without disturbing the whole. It is a word-picture in a similar sense in which the picture of a painted picture worked in our portrayal of yearning: It helps indicate the nature of the object meant without attempting to describe its nature.[10] One reason for proceeding in this way may be this, that no attempt at description would be helpful, and all would be subject to Wittgenstein's criticism.

Notes

1. The abbreviation 'L.L.' [= loose-leaf folder] was used by Wittgenstein to mark those passages in his manuscripts that he wished to transfer to this folder. Cf. Schulte, 1993, p. 6.
2. Manuscript numbers are given in accordance with von Wright's catalogue of Wittgenstein's Nachlass (reprinted in PO, pp. 480–506). Quotations are given and translated from the Bergen Electronic Edition (= BEE) of Wittgenstein's Nachlass.
3. There is something a little odd about the German sentence, which suggests that the soul craves for the desire. Presumably, what is meant is that it is the object of the desire which is craved for. (The German original runs as follows: 'Und man könnte auf diese Weise freilich auch darstellen, wie ein Mensch in seiner Miene dem Wunsch keinerlei Ausdruck gibt, und doch seine Seele nach ihm verlangt'.)
4. Lighter colours and broken outlines are standard ways of indicating that a figure is the soul. But often (e.g. in scenes of death) the depicted context makes it unmistakably clear which figure is the body, and which the soul.

5. The typescript versions of the two previous remarks can be found in RPP I, §§275–6; the manuscript remarks on hope and fear are published in BEE, but not in printed form.
6. In his first manuscript remark of the following day (21.8.46; cf. RPP I, §282), Wittgenstein returns to the topic of how to understand the difference between 'psychological phenomena' which can be pictorially represented without resorting to such means (e.g. writing) and those that cannot be depicted without resorting to them. This remark is followed by observations on differences in the 'nature' of various 'psychological phenomena'. These observations are particularly interesting because they effect a connection with Wittgenstein's reflections on the natural history of man (cf. PI, §25; RPP II, §18 = MS 135, pp. 83v–84r [10.12.47]; the manuscript entries of this day are more comprehensive and more instructive than the material printed in RPP II).
7. Cf. PI, §291 for a contrast between pictures ('descriptions') as 'instruments for particular uses' and 'idle' pictures, which 'seem simply to depict how a thing looks, what it is like'.
8. This modification is relatively late. Both the manuscript (August 1946) and the typescript (autumn 1947) have 'Der Mensch ist das beste Bild der menschlichen Seele'. It is only in the spring of 1949 that Wittgenstein introduces the modified version we know from the *Investigations*: 'Der menschliche Körper ist das beste Bild der menschlichen Seele.'
9. Cf. Malcolm (1963, p. 118). In this passage, Malcolm discusses Wittgenstein's tale of soulless slaves, which is relevant in this context. But by bringing in an 'as if' into our (the slaveholders') way of regarding the slaves, he may be spoiling the, or one, point of this tale.
10. To adapt the words I used above: context, tradition and conventions of using and reading emblematic expressions join forces to make us see that such expressions are used and mentioned at the same time.

References

S. Cavell (1979) *The Claim of Reason: Wittgenstein, Scepticism, Morality, and Tragedy* (New York and Oxford: Oxford University Press).

N. Malcolm (1963) 'Wittgenstein's *Philosophical Investigations (1954)*' in N. Malcolm *Knowledge and Certainty* (London: Cornell University Press).

J. Schulte (1993) *Experience and Expression: Wittgenstein's Philosophy of Psychology* (Oxford: Clarendon Press).

4

Aspect Seeing in Wittgenstein and in Psychology

Nicole Hausen and Michel ter Hark

This Paper explores the relevance of conceptual analysis to psychology by considering how Wittgenstein's investigations relate to influential psychological work of his own day. In particular, we focus on Wittgenstein's engagement with the writings of prominent Gestalt psychologist Wolfgang Köhler (1887–1967), and we discuss ways in which Wittgenstein's studies of aspect seeing yield critiques of Köhler's theory of perception. Since Wittgenstein directly addresses Köhler's ideas, our Paper thereby shows how Wittgenstein's own words reveal conceptual weaknesses in a psychological theory.

The text proceeds as follows: Section 4.1 provides a short survey of Wittgenstein's interest in aspect seeing, the phenomenon in which, for instance, an ambiguous figure is seen as either one thing or another. We give examples of figures that Wittgenstein discusses, including some that appear in Gestalt psychology publications as well, and we mention issues surrounding aspect seeing that received Wittgenstein's attention. In addition, Section 4.1 has a methodological component in which we briefly explicate Wittgenstein's view that philosophical investigation is 'grammatical' (that is, conceptual) in nature and illustrate how this view pertains to his research on aspect seeing. Section 4.2 describes main tenets of Köhler's theory of perception and how this theory accounts for the phenomenon of aspect seeing. In Section 4.3, we utilize remarks by Wittgenstein in order to develop conceptual critiques of Köhler. Points highlighted are Köhler's notion of organization of the visual field and the voluntariness of aspect seeing. Finally, Section 4.4 offers concluding comments about what the conceptual analysis presented in this Paper implies (and does not imply) for Köhler's theory, and about what general lessons can be taken away from this one historical case study.

4.1 Wittgenstein and aspect seeing

This section contains an introduction to Wittgenstein's investigations of aspect seeing. We also discuss the philosophical method that underlies Wittgenstein's studies. This method is central (in Section 4.3) to our interpretations of Wittgenstein's remarks and to their application to Köhler's work.

4.1.1 An overview

Aspect seeing was a topic that intrigued Wittgenstein at points throughout his career, but especially between 1946 and 1949. He deals with aspect seeing in manuscripts and typescripts from 1946–1948 that have now been published as *Remarks on the Philosophy of Psychology* I and II, and it is a prominent subject in manuscripts from 1948–1949 that have appeared as *Last Writings on the Philosophy of Psychology* I. Aspect seeing is also discussed in considerable length in the selection of remarks that has been printed under the title *Part II* of *Philosophical Investigations* (and recently reprinted as *Philosophy of Psychology – A Fragment*). Wittgenstein culled out those remarks in 1949 from previous work, and they date back to 1946, with the majority coming from 1948 and 1949. The students' course notes that are recorded in *Wittgenstein's Lectures on Philosophical Psychology 1946–47* show that Wittgenstein covered aspect seeing in his lectures during that time period as well. Yet even before 1946, Wittgenstein's interest in aspect seeing is evident. The *Brown Book*, which Wittgenstein dictated to two pupils in the academic year 1934–1935, contains thoughts regarding aspect seeing; and indeed, an awareness of issues surrounding aspect seeing is already present in the *Tractatus Logico-Philosophicus*, published in 1922.

In the *Tractatus,* Wittgenstein's concern with aspect seeing is fleeting, consisting of a reference to the so-called Necker cube (Figure 4.1). As Wittgenstein indicates, there are two ways of seeing this figure. Namely, it can be seen as a cube oriented with vertices *a* in front, or it can be seen as a cube oriented with vertices *b* in front. Wittgenstein does not go much farther than this, adding just a short comment about how the phenomenon fits within the philosophical framework of the *Tractatus*. But although Wittgenstein only touches on aspect seeing in the *Tractatus*, the Necker cube is indicative of the kinds of ambiguous figures that pervade Wittgenstein's later, intensive studies of aspect seeing.

Perhaps the most famous figure used by Wittgenstein is the duck–rabbit (Figure 4.2), which can be seen as a drawing of a duck or as a drawing of a rabbit. The triangle (Figure 4.3) can be seen, for instance, as sitting on

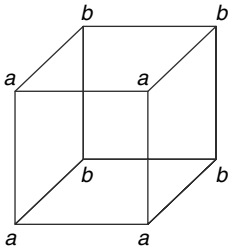


Figure 4.1 The Necker cube
Source: See TLP, 5.5423.

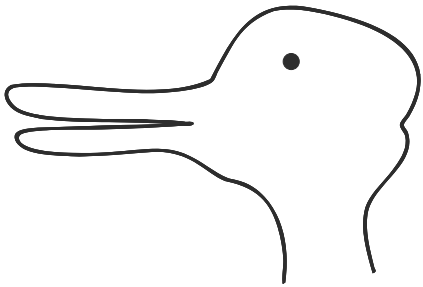


Figure 4.2 The duck–rabbit
Source: See PPF, §118.
Note: The duck–rabbit figure also appears in LPP, p. 104; RPP I, §70.

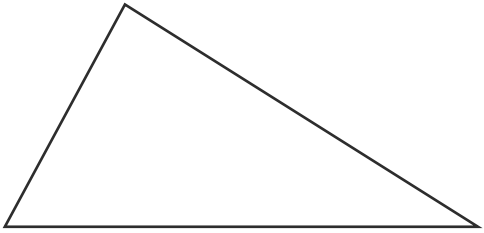


Figure 4.3 The triangle
Source: See PPF, §162.
Note: The triangle figure also appears in LPP, pp. 100, 229, 329; LW I, §605; PPF, §205.

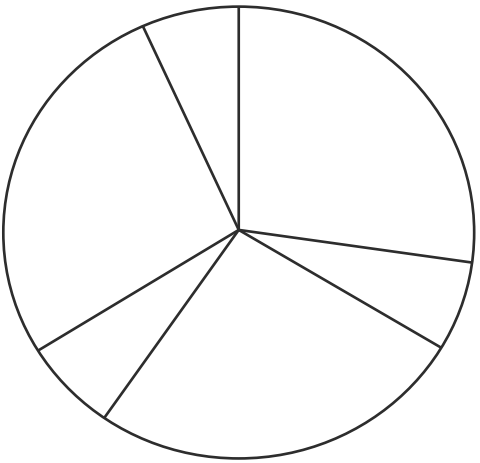


Figure 4.4 The 'pie'
Source: See RPP I, §1117.
Note: The 'pie' figure also appears in LPP, pp. 114, 343, 345.

its base or as lying on its side or as hanging, or even as a triangular hole. The lines in the 'pie' figure (Figure 4.4) can be seen as three pairs of lines that bound three narrow arms with large gaps in between; or, grouping the lines into different pairs, they can be seen as bounding three wide arms with small gaps in between. In a similar vein, the drawing in Figure 4.5 can be seen as two unrelated ovals with a horizontal line through them; or instead, the parts of the figure can be seen as each contributing a component to form a '4' that is positioned between the two ovals. The double cross (Figure 4.6) can be seen as a black cross on a white ground or as a white cross on a black ground.

Several of the figures shown here express Wittgenstein's familiarity with salient Gestalt psychology literature. For Köhler discusses the pie figure and a variation of the 4 figure (1947, pp. 171–2, 185–6), whereas versions of the double cross are presented by Kurt Koffka (1886–1941), another eminent Gestalt psychologist (1935, pp. 83, 191).[1] In addition, Wittgenstein occasionally refers to Köhler explicitly in his written remarks, and he mentions him by name throughout his lectures in 1946–1947. Wittgenstein is furthermore reported to have begun many of those lectures by reading short selections from Köhler's *Gestalt Psychology* (Monk, 1990, p. 509). Given these connections, it is not surprising that some of Wittgenstein's work on aspect seeing bears directly on issues in Gestalt psychology.

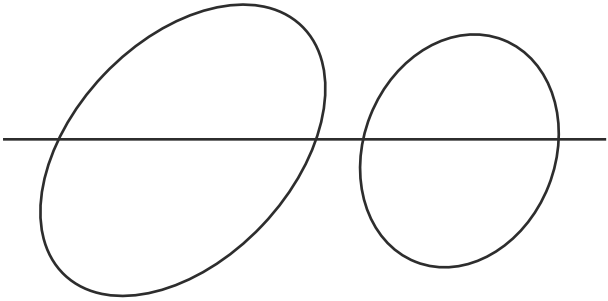


Figure 4.5 The '4'
Source: See RPP II, §41.
Note: A variation of the '4' figure also appears in RPP I, §982.

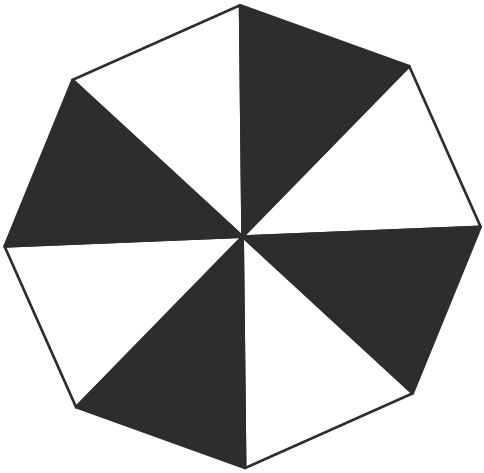


Figure 4.6 The double cross
Source: See RPP I, §971.
Note: The double cross figure also appears in PPF, §212; RPP I, §1017.

Wittgenstein's thoughts concerning claims made by Gestalt psychology are intermixed with his other investigations of aspect seeing, and overall, his studies of aspect seeing have a wide scope. For example: Wittgenstein looks at the relation of aspect seeing to 'regular' seeing and to interpretation and thought. He distinguishes different kinds of aspects, such as 'conceptual' aspects (aspects of the triangle, for

instance) and 'organizational' aspects (aspects of the pie, for instance). He studies changes in aspect, like seeing Figure 4.2 as a picture of a duck and then seeing it as a picture of a rabbit. Correspondingly, he addresses the puzzling character of aspect seeing, in which something changes, and yet nothing changes. But despite the diversity of subjects involved, Wittgenstein's underlying methodological approach to aspect seeing is invariant.[2] In particular, his investigations are conceptual (that is, grammatical), rather than empirical. It is therefore conceptual analysis that we will use in order to critique Köhler's ideas. The notion of 'conceptual analysis' requires some elucidation, however. It is also useful to give an indication of how it can be expected to have a bearing on scientific work. It is these two topics that we will address in the remainder of this section.

4.1.2 Philosophical method

Unlike scientific work in psychology, Wittgenstein's philosophy of psychology does not posit and test hypotheses or develop explanatory accounts of the phenomena in question. Wittgenstein instead studies psychological *concepts*, such as the (everyday) concepts of seeing, of thought and of emotion. Moreover, he says that these concepts are inherent in our everyday use of language. His philosophical investigations therefore involve the description of what he calls 'grammar', that is, the ways that words are used. By describing the grammar of our words, Wittgenstein is describing our concepts. Or, to formulate the idea differently, he is describing what is meaningful, or *makes sense*, in our language.

Wittgenstein's viewpoint regarding the philosophical study of seeing, for example, is summed up in the following remark, which alludes to Köhler's theory of perception and his claim that we see a glance of the eye just as we see colour (cf. Köhler, 1947, pp. 232–3):

> 'When you get away from your physiological prejudices, you'll find nothing in the fact that the glance of the eye can be seen.' Certainly I too say that I see the glance that you throw someone else. And if someone wanted to correct me and say I don't really *see* it, I should hold this to be a piece of stupidity.
>
> On the other hand I have not *admitted* anything with my way of putting it, and I contradict anyone who tells me I see the eye's glance 'just as' I see its form and colour.
>
> For 'naïve language', that's to say our naïf, normal, way of expressing ourselves, does *not* contain any theory of seeing – it shews you, not

> any theory, but only a *concept* of seeing. (RPP I, §1101, emphasis original)

The comments in RPP I, §1101 reveal in Wittgenstein's own words that he is not attempting to give an explanatory account of seeing. He is instead interested in the ordinary ways we use the word 'see'. And in this ordinary sense, it is perfectly fine to say that we see a glance. But according to Wittgenstein, this means that it is part of our *concept* of seeing that we *see* a *glance*. It does not constitute a hypothesis (in a theory of perception) that tells us what counts as 'genuine seeing'.

To further understand what it means to investigate seeing grammatically, consider this remark:

> Two uses of the word 'see'.
>
> The one: 'What do you see there?' – 'I see *this*' (and then a description, a drawing, a copy). The other: 'I see a likeness in these two faces' – let the man to whom I tell this be seeing the faces as clearly as I do myself.
>
> What is important is the categorial difference between the two 'objects' of sight. (PPF, §111, emphasis original)

Here, Wittgenstein is thinking about two different ways in which the word 'see' can be used legitimately, and these two different uses involve two grammatically-different kinds of 'objects' of sight. The two uses are captured in sentences such as (1) 'I see a face' and (2) 'I see a likeness between the two faces'. In the case of (1), for example: Saying what one sees involves saying something like 'I see a face', and then continuing with a description of it, such as 'It is an oval shape, and the eyes are brown, and it looks like *this* [drawing a face]'. It does not make sense for someone to say, 'I see a face, but I can give no further description of it'. Also, when a (sighted) speaker A is looking at a face that is plainly in front of him, it does not make sense for A to say, 'I don't see a face'.[3] These points contrast with (2). With respect to (2), it makes sense to say 'I see a likeness, but I can't describe it' (cf. RPP II, §551). Similarly, when A is looking at the two faces, it does make sense for him to say 'I see two faces, but I don't see a likeness' (cf. PPF, §112). This grammatical analysis shows that our everyday concept of seeing applies to both faces and likenesses (that is, it makes sense to say both that we see a face and that we see a likeness), yet the word 'see' is used in different ways with these different 'objects' of sight.

The question remains as to how such conceptual analysis can have relevance to scientific research. But as Maxwell Bennett and Peter Hacker

(2003) have emphasized, conceptual questions precede empirical ones. Hence, for example, before one can ask whether a particular scientific statement (like the statement that we see a glance just as we see colour) is true or is supported by experimental results, one must ascertain whether the claim even *makes sense*, given the concepts at issue (in this case, our ordinary concept of seeing). It is this latter task that conceptual analysis can help to achieve.[4] In Section 4.3 below, we concentrate on a similar application of conceptual analysis to scientific work. Specifically, we use conceptual analysis to investigate Köhler's notion of 'organization' and also the relation of that notion to the voluntariness of changes in aspect. Our aim is to thereby illustrate how Köhler's lack of attention to conceptual details ultimately undermines the value of his proposal. Prior to that study, however, pertinent claims made by Köhler are outlined next in Section 4.2.

4.2 Köhler and Gestalt psychology

One instructive way to view Gestalt psychology is in contrast to the framework that Köhler calls 'Introspectionism'. Köhler himself in fact devotes an early chapter of *Gestalt Psychology* to an exposition and critique of Introspectionism. In doing so, he focuses not on any one psychological school but rather on certain assumptions about perception that can underlie various psychological traditions.[5] As Köhler describes it, Introspectionism makes a crucial distinction between sensations and perceptions. Sensations for an Introspectionist are 'bare sensory material as such', whereas perceptions couple effects of learning to this sensory material. Moreover, according to Introspectionists, genuine seeing involves sensations only (1947, pp. 68–9). When we look at the desk in front of us, we thus *perceive* a grey object because our previous interactions with objects (and in particular, our interactions with desks) impart meaning to the grey patch of colour that we would *see*. That we do not seem to see simply a grey patch of colour is due to the effects of learning (arising from, for example, our interactions with objects), and it is a task of the trained Introspectionist to uncover the visual sensations that learning has masked (cf. Köhler, 1947, pp. 69, 78–80). Köhler concludes from this that, in our everyday visual experiences, 'little is left that would be called a true sensory fact by the Introspectionist' (1947, p. 83).

Köhler rejects the Introspectionist account of perception. His alternative claim is that our visual field has an 'organization', and this organization is a sensory (specifically, a visual) fact, just like colour. According to Köhler, it is in virtue of organization that 'the contents of particular

areas [in the visual field] "belong together" as circumscribed units from which their surroundings are excluded' (1947, pp. 137, 139). Köhler comments that sometimes the segregated entities, or *Gestalten*, that are formed in visual organization correspond to physical units (such as in the case of objects like desks), and sometimes they do not (such as in the case of star constellations). Conversely, sometimes physical units do not correspond to Gestalten in the visual field, as in the case of camouflaged objects (1947, pp. 156–7, 160).

Because it is a sensory fact, organization is not the result of learning. Yet, Köhler admits that learning does often add meaning to Gestalten. For instance, we may learn that a certain segregated unit before us is called a 'desk' and that it is useful as a workspace. (This is analogous, Köhler says, to learning that a certain colour is called 'green' and that green is a symbol of hope.) Köhler maintains, however, that the segregated wholes are given first as visual facts, and *then* we associate meanings with them (1947, pp. 138–9, 192). Köhler stresses also that when sources outside the organism stimulate the retina, the resulting 'mosaic' on the retina is not itself already organized into Gestalten. Instead, the nervous system responds to the retinal stimulation, and various Gestalten in the visual field can thereby result (1947, pp. 160–2, 180–2). Gestalt psychologists thus postulate principles of organization that govern the formation of Gestalten. For example, certain *kinds* of Gestalten tend to form, such as simple, closed areas instead of irregular and open units (1947, pp. 144–5). This principle seems to give a reason why Figure 4.5 is typically seen as two ovals and one line segment, rather than as a 4 that 'splits up' the line segment into three and 'takes away' portions of each oval (cf. Köhler, 1947, pp. 185–6).

Another important facet of Gestalt psychology is that it adheres to a thesis of psychophysical isomorphism, which Köhler describes as 'the thesis that our experiences and the processes which underlie these experiences have the same structure' (1947, p. 344). In other words, all psychological experiences are correlated with events in the central nervous system, and the corresponding psychological and physical events are structurally similar (Köhler, 1940, pp. 47–8). Köhler advocates studying physiological events in addition to making qualitative and quantitative observations of behaviour, claiming that 'purely psychological research is not likely to yield a systematic theory of mental facts' (1940, p. 46; 1947, chapter II). As an example, he cites the principles of organization mentioned above. Although these principles help predict which organizations will occur in the visual field, they do not tell us why *these* principles hold and not others, or in what way

the particular factors set out in the principles favour the formation of Gestalten (1940, p. 44).

Köhler's characterisation of perception implies that we see an aspect such as the black cross of Figure 4.6 in virtue of the organization of our visual field. Moreover, since organization is a visual fact, we genuinely see the cross-shaped Gestalt just as we see the black and white. Hence, what Wittgenstein calls a 'change of aspect' in this case is a sensory change. When we see a black cross-shaped Gestalt and then see a white cross-shaped Gestalt, visual facts have changed. In Köhler's words, there is a 'transformation in [the] visual field' (1947, p. 185). Köhler also goes beyond this and, in striving for an understanding of the physiological basis of perception, offers a physiological hypothesis related to change in aspect. He discusses experiments he had done with subjects who viewed Figure 4.4 and were told to 'keep' whatever aspect they saw at a given time. The results from some subjects show that the longer the subject looks at the figure, the more rapidly the two aspects alternate. Köhler suggests that this is due to electric currents in the nervous system that alter the tissue through which they pass. Specifically, the currents alter the tissue in a way that impedes their own passage. Thus, after seeing one aspect of the pie figure for a certain amount of time, it becomes more difficult to see that aspect, and the aspect consequently changes. This then occurs for the new aspect, resulting in a more rapid alteration of aspects (1940, pp. 67–74; 1947, pp. 171–2). Köhler also proposes the following: The idea that there is a process occurring that changes the medium in which it occurs applies not just to seeing aspects of an ambiguous figure like the pie, but to perception more generally. The distinctive condition with respect to seeing aspects is that a change in aspect takes place when the physiological change in the medium reaches a certain critical level. Prolonged viewing of ordinary (that is, unambiguous) figures should therefore be expected to result in at least some transformation in the visual field as well (1940, pp. 82–3).

4.3 Conceptual studies of aspect seeing and their application to Gestalt psychology

We have described Wittgenstein's philosophical work and Köhler's scientific work in the previous sections. The present section seeks to show, via two concrete examples, how Wittgenstein's observations regarding the use of language can point to problems with the Gestalt approach embraced by Köhler. The first example deals exclusively with the notion of organization that is part of Köhler's theory of perception.

The second example extends our critique of Köhler's notion of organization while centring on the (conceptual) voluntariness of aspect seeing. But along with these criticisms, we also provide a 'positive' interpretation of Wittgenstein's remarks regarding Köhler's use of the term 'organization'.

4.3.1 Köhler's 'organization' of the visual field

Authors who discuss Wittgenstein's comments concerning Köhler's notion of organization typically focus on the negative critique.[6] We, too, will say something about Wittgenstein's negative and more familiar criticisms of Köhler. Specifically, we will have something to add to the discussion in terms of the grammatical genesis of certain problems with Köhler's theory. Yet, from Wittgenstein's remarks about organization in the context of aspect seeing, one can also glean an alternative (positive) description of the use of this concept. We will thus present this alternative description of the use of the word 'organization' as well. Both our negative and positive arguments rely on a subtle and rarely-discussed distinction that Wittgenstein makes between the transitive and intransitive use of terms, so we turn now to that distinction.[7]

Wittgenstein develops the transitive/intransitive distinction in the *Brown Book*, immediately prior to a discussion of aspect seeing. His example is the word 'particular', as when we say, 'The face has a particular expression'. Here is how he introduces the difference between the transitive and intransitive use of the word 'particular' as applied to familiar items of various sorts:

> Now the use of the word 'particular' is apt to produce a kind of delusion and roughly speaking this delusion is produced by the double usage of this word. On the one hand, we may say, it is used preliminary to a specification, description, comparison; on the other hand, as what one might describe as an emphasis. The first usage I shall call the transitive one, the second the intransitive one. Thus, on the one hand I say 'This face gives me a particular impression which I can't describe'. The latter sentence may mean something like: 'This face gives me a strong impression'. These examples would perhaps be more striking if we substituted the word 'peculiar' for 'particular', for the same comments apply to 'peculiar'. If I say 'This soap has a peculiar smell: it is the kind we used as children', the word 'peculiar' may be used merely as an introduction to the comparison which follows it, as though I said 'I'll tell you what this soap smells like: ...'. If, on the other hand, I say 'This soap has a *peculiar* smell!' or 'It has a most

> peculiar smell', 'peculiar' here stands for some such expression as 'out of the ordinary', 'uncommon', 'striking'. (BBB, p. 158, emphasis original)

So, in the transitive case, the word 'particular' (or 'peculiar') is used as a precursor to a further specification. To the question 'Peculiar in what way?' an answer can be given that explains this way in different words. In the intransitive case, however, the word 'particular' is used for emphasis, and hence there is no further specification or comparison to be made.

Transitive and intransitive uses of words are not always easy to tell apart, however. This is especially true when the sentences in question involve what Wittgenstein calls a 'reflexive construction' or a 'reflexive expression' (BBB, pp. 159–61). The use of words in a reflexive construction is intransitive yet *appears* to be a special case of a transitive use (namely, the reflexive constructions appear to be comparing something with itself or describing something by appealing to the thing itself). The important feature of reflexive constructions is that the sentences can be, as Wittgenstein says, 'straightened out'. What he means by this is that the sentences seem to involve a comparison or description that loops from an object back to itself. But when the sentences are 'straightened out', we see that there is no loop. Rather, the sentences involve only an intransitive use; that is, they involve emphasis, not comparison or description. For instance, Wittgenstein says that 'That's that' is a reflexive expression. Although 'That's that' appears to compare a thing to itself, it can be straightened out as 'That's settled' and in fact is used to emphasize the finality of the situation.

Looking at another of Wittgenstein's examples will help to clarify these ideas. Wittgenstein supposes that he has been observing someone as that person sits in the room. He then comments:

> Now if I wished to draw him as he sat there, and was contemplating, studying, his attitude, I should while doing so be inclined to say and repeat to myself 'He has a particular way of sitting'. But the answer to the question 'What way?' would be 'Well, *this* way', and perhaps one would give it by drawing the characteristic outlines of his attitude. On the other hand, my phrase 'He has a particular way … ', might just have to be translated into 'I'm contemplating his attitude'. Putting it in this form we have, as it were, straightened out the proposition; whereas in its first form its meaning seems to describe a loop, that is to say, the word 'particular' here seems to be used transitively and, more

> particularly, reflexively, i.e., we are regarding its use as a special case of the transitive use. We are inclined to answer the question 'What way do you mean?' by '*This* way', instead of answering: 'I didn't refer to any particular feature; I was just contemplating his position'. (BBB, p. 160, emphasis original)

In other words, upon reflection (and upon straightening out the sentence 'He has a particular way of sitting – *this* way [pointing to the sitting man, for example]'), we see that we are not intending in this context to point out something *about* the man's way of sitting. We are instead using the sentence to draw attention to the man's sitting and to emphasize that we are having a certain experience with respect to it, like that of contemplating it or trying to draw it (BBB, p. 160). That is, we are using 'particular' intransitively. This contrasts with a transitive use of 'particular' in 'He has a particular way of sitting', such as when we answer 'What way?' by saying, 'He is sitting with his toes pointed up'.

Wittgenstein's objective in discussing these distinctions is to point out that confusions can arise if transitive and intransitive uses are not properly distinguished. Although Wittgenstein does not discuss Köhler within the context of transitive and intransitive use, we will argue here that Köhler's notion of organization falls into the transitive/intransitive trap. Specifically, it looks as if Köhler is using the term 'organization' transitively when he speaks about the organization of the visual field. But actually what is involved is an intransitive use. By failing to recognize this, Köhler also fails to see certain problems that afflict his theory of perceptual organization.

To begin, it is necessary to outline what some of the relevant weaknesses in Köhler's theory are. As explained in Section 4.2, Köhler defines organization of the visual field as a sensory fact in addition to colour. So, when we experience a change in aspect of (for example) the pie figure, there is a change in the sensory facts, namely, the organization of our visual field changes. But as Wittgenstein says, '"The organization of the visual image changes" has not the same kind of application as: "The organization of this company is changing." *Here* I can describe *how it is*, if the organization of our company changes' (RPP I, §536, emphasis original).[8] That is, a company's organization may be described by a flowchart that shows the company's hierarchy and structure. It makes sense to ask, '*How* did the organization change?', and the response could involve pointing to changes in the flowchart. But there is no comparable way to describe the organization of a visual field. We might, as Wittgenstein suggests, represent our visual impressions by means of drawings. Such drawings

would reflect a change in colour. Yet, these drawings will show no change when there is a change in aspect – they will be the same before and after the theorized change in organization takes place (LW I, §439/PPF, §131). Wittgenstein indicates that therefore Köhler must be taking an 'organized visual field' to be some sort of inner object (LW I, §443/PPF, §134). According to Wittgenstein, Köhler would say that it is this inner mental object that shows the change of organization.

Commentators such as Malcolm Budd and Michel ter Hark, who have also discussed these issues, conclude that Köhler's notion of organization is, for instance, 'unilluminating' (Budd, 1989, p. 85) and 'mystifying' (ter Hark, 1990, pp. 175–6). Budd bases his charge on Köhler's inability to account for *how it is* when the organization of the visual impression changes. Ter Hark maintains that by involving unspecified inner objects, Köhler has just transferred the problem to a new domain. For we are left without any information about the objects that are supposed to render the notion of organization intelligible.

What we propose is that the transitive/intransitive distinction can be applied to the word 'organization', and this sheds new light on Köhler's problems. With respect to a sentence like 'The organization of the company has changed', it is possible to go on to specify the change. Likewise, the relevant change can be specified for a sentence such as 'The colour of the sky has changed'. The sentence 'The organization of my visual field has changed' *seems* similar. When one tries to answer the question 'How has the organization of your visual field changed?', however, the most one can do is refer to the inner mental objects and say, 'Like *this*'. And this response is analogous to describing a man's particular way of sitting by simply drawing a picture of him sitting.[9] In other words, the response is not truly an informative further specification at all, and the use of 'organization' in 'The organization of my visual field has changed' is not transitive. Rather, the sentence involves a reflexive construction, and the use of 'organization' is intransitive.

In making this argument, Wittgenstein's comments in RPP I, §1118 (which immediately follow a remark about Köhler and the pie figure) also are relevant to discuss. Wittgenstein notes:

> Indeed, you may well say: what belongs to the description of what you see, of your visual impression, is not merely what the copy shews but also the claim, e.g. to see this 'solid', this other 'as intervening space'. Here it all depends on *what we want to know* when we ask someone what he sees. (RPP I, §1118, emphasis original)

In fact, a central idea in Wittgenstein's analysis of aspect seeing is that in everyday contexts the change in what is seen *is* adequately described by (for instance) pointing to part of the pie and saying 'I used to see *this* part of the figure as intervening space, and now I see it as solid'. For example, if Wittgenstein was looking at the pie figure and wanted to describe a change in what he sees, he could say 'I now see the narrow sectors as solid'. The situation is different in Köhler's case, however. Suppose that Köhler would suggest that 'I now see the narrow sectors as solid' describes the change in organization of the viewer's visual field. That is, suppose that Köhler were to suggest that a (transitive) answer to 'How has the organization of your visual field changed?' is 'I now see the narrow sectors as solid'. In this case, the answer is not sufficient. The reason why it is insufficient is because Köhler needs the answer to provide more than just a description of the change in what is seen. For he intends to *explain* change in what is seen by appeal to change in organization of the visual field. Yet, saying that I now see the narrow sectors as solid (this is how the organization has changed) does not *explain why* I now see the narrow sectors as solid (this is what I now see). In other words, Köhler would be claiming, in effect, 'I now see the narrow sectors as solid *because* I now see the narrow sectors as solid', which clearly does not provide an informative explanation. The suggested response (that the change in organization is described by, for instance, 'I now see the narrow sectors as solid') would therefore undercut the explanatory power that Köhler wants his notion of organization to have. Correspondingly, it does not provide him with a transitive use of the word 'organization'.

Viewing Köhler's situation from the transitive/intransitive perspective hence exposes yet another way that his notion of organization is unilluminating and mystifying. When introducing organization as a sensory fact, Köhler apparently assumes that his notion will have a transitive use similar to that of our concepts of (ordinary) organization and of colour. But, upon inspection, we see that Köhler's notion lacks any transitive use at all. Even if Köhler would, when pressed, identify an organized visual field with some sort of inner mental object, this does not help him further specify what he means by 'organization'. He is thus using a term that, in an important way, lacks content.

On that note, we turn to Wittgenstein's alternative, positive description of the use of the word 'organization' with respect to aspect seeing. We have argued above that there is not a transitive use of Köhler's 'organization' and that instead it is used as part of reflexive constructions (that is, intransitively). But then it should be possible to 'straighten out' a sentence such as 'The organization of my visual field has changed'.

This 'straightening out' is, we suggest, the essence of Wittgenstein's positive description of Köhler's notion of organization.

Consider the following comments:

> If someone says: 'I am talking of a visual phenomenon, in which the visual picture, that is its organization, does change, although shapes and colours remain the same' – then I may answer him: 'I know what you are talking about: I *too should like to say* what you say.' – So I am not saying 'Yes, the phenomenon we are both talking about is actually a change of organization …' but rather 'Yes, this talk of the change of organization etc. is an expression of the experience which I mean too.' (RPP I, §534, emphasis original)
>
> 'The organization of the visual image changes.' – 'Yes, that's what I'd like to say too.'
>
> This is analogous to the case of someone saying 'Everything around me strikes me as unreal' – and someone else replies: 'Yes, I know this phenomenon. That's just how I'd put it myself.' (RPP I, §535)

Wittgenstein's idea in these remarks is that a sentence like 'The organization of my visual field has changed' can be useful even if it is not used as a precursor to a specification of *how* the organization has changed. In particular, it can be used to express (and thereby emphasize) an experience that one has had. So, 'The organization of my visual field has changed' is straightened out as 'I'm having an experience that I want to express by saying "The organization of my visual field has changed"'. Moreover, this experience need not be further explicated in order for the sentence to be meaningful. As Wittgenstein says regarding the feeling of everything being unreal, 'And how do I know that another has felt what I have? Because he uses the same words as I find appropriate' (RPP I, §125). Hence, Köhler's term 'organization' *does* have value. But this value lies with speakers' common use of the word 'organization' in connection with experiences of changes in aspect, and not with a specification of the change in organization itself.[10] By speaking of the organization of our visual field, we are (intransitively) emphasizing an experience rather than (transitively) describing the visual field.

4.3.2 The voluntariness of aspect seeing

Having covered the first of our two examples of conceptual criticism of Köhler, we move on to the second. This second example builds on Wittgenstein's comments in RPP I, §971 (emphasis original): 'What Köhler does not deal with is the fact that one may *look at* [the double

cross] in this way or that, that the aspect is, at least to a certain degree, subject to the will'. Now, although Wittgenstein's remarks about being subject to the will are commonly acknowledged to be conceptual points, application to Köhler's work is lacking.[11] In this part of the Paper, we will therefore show grammatically how this idea from Wittgenstein yields another critique of Köhler's theory, and in particular, a critique of his notion of organization as a sensory fact like colour.

In order to carry out our argument, we first need to explain in grammatical terms what Wittgenstein means when he says that something is subject to the will (that is, voluntary). Consider the following remark, which does not involve aspect seeing but is very clear with respect to being subject to the will:

> If I say that imaging is subject to the will that does not mean that it is, as it were, a voluntary movement, as opposed to an involuntary one. For the same movement of the arm which is now voluntary might also be involuntary. – I mean: it makes *sense* to order someone to 'Imagine that', or again: 'Don't imagine that.' (RPP II, §83, emphasis original)

Wittgenstein's thinking is this: There is a sense, an empirical sense, in which certain actions are (in)voluntary. For example, if a man chooses to move his arm, then the subsequent movement of the arm is voluntary. Yet, a similar movement of the arm might be involuntary, such as if he is startled, and his arm moves automatically as part of a startle-response, or if another person moves the arm against his will. In this empirical sense, the claim that imagining is at all times voluntary is false, since images sometimes occur automatically, or we try to form an image of something and cannot, or we have images that we want to get rid of but cannot. But grammatically or conceptually, imagining *is* always voluntary because it always *makes sense* to issue certain commands, such as 'Imagine a tree' or 'Don't imagine a tree', regardless of whether the commands are actually carried out successfully.

This helps give grammatical context to two remarks about voluntariness that do involve aspect seeing:

> Seeing an aspect and imagining are subject to the will. There is such an order as 'Imagine *this*!', and also, 'Now see the figure like *this*!'; but not 'Now see this leaf green!'. (PPF, §256, emphasis original)
>
> An aspect is subject to the will. If something appears blue to me, I cannot see it red, and it makes no sense to say 'See it red'; whereas it

> does make sense to say 'See it as…'. And that the aspect is voluntary (at least to a certain extent) seems to be essential to it, as it is essential to imaging that *it* is voluntary. (RPP I, §899, emphasis original)

In other words, a command like 'See the figure as a black cross' makes sense, whereas a command like 'See this leaf green' does not. Hence, aspect seeing is conceptually voluntary or subject to the will, and seeing colours is not.[12]

Next, it is relevant to reflect on the case of physical objects (and, more precisely, the Gestalten with which they are correlated). At the very beginning of his discussion of organization of the visual field, Köhler says:

> On the desk before me I find quite a number of circumscribed units or things: a piece of paper, a pencil, an eraser, a cigarette, and so forth. The existence of these visual things involves two factors. What is included in a thing becomes a unit, and this unit is segregated from its surroundings. In order to satisfy myself that this is more than a verbal affair, I may try to form other units in which parts of a visual thing and parts of its environment are put together. In some cases such an attempt will end in complete failure. In others, in which I am more successful, the result is so strange that, as a result, the original organization appears only the more convincing as a visual fact. (Köhler, 1947, pp. 137–8)

Köhler apparently intends this as empirical support of his claim that organization is a sensory fact. That is, formation of visual units (Gestalten) is strictly a product of neural function and is not dependent on any cognitive processes (such as trying to see different visual units). We suggest, however, that Köhler may be over-reaching in his attempt to defend his position. Arguably, it does not even *make sense* to 'put parts of a visual thing together with its environment' when the visual thing corresponds to a physical object. If someone were ordered to 'See the two units as one unit' or 'See the unit and the ground behind it as one unit', it would be unclear what these orders mean. At most, we might interpret them as orders about seeing aspects, like 'Imagine that the two units are glued together' (which would be similar to seeing the triangle figure as hanging) or 'See the two units as belonging together' (which would be similar to seeing certain stars as belonging together). However, Köhler wants seeing segregated units to be like seeing colour. So, it seems that an analogous command to 'See the two units as one unit' would be 'See

the red surface as blue'; and here, it is obvious that neither command makes sense. Similarly, a more basic command like 'See *this* [object] as a unit' seems analogous to 'See this leaf as green'.

Hence, Köhler faces a dilemma. Either seeing segregated units that correspond to physical objects is conceptually voluntary, like seeing aspects, or it is conceptually involuntary, like seeing colour. In the former case, this implies that seeing Gestalten in general is *not* like seeing colour. In the latter case, it turns out that there are conceptual differences between different kinds of Gestalten (namely, seeing Gestalten that correspond to physical objects is conceptually involuntary, while seeing Gestalten that are aspects is conceptually voluntary). This means that either Köhler is ignoring conceptual differences between an organized visual field and colour (because seeing Gestalten differs from seeing colour) or he is ignoring conceptual differences between different 'objects' of sight (because seeing segregated units that correspond to physical objects differs from seeing aspects).[13]

Either way, there is a problem for Köhler's theory. Köhler wants to treat everything in a uniform way and explain all the Gestalten we see by the notion of an organized visual field that is on a par with colour. But due to the conceptual factors we have discussed above, either the parallel between organization and colour breaks down entirely, or organization cannot account for seeing aspects *just as* it accounts for seeing Gestalten that correspond to physical objects. That is, there is more conceptual diversity in perception than Köhler's theory based on an organized visual field allows.

This also raises issues for his hypothesis regarding electric currents that was derived from experiments involving changes of aspect. In particular, Köhler uses the results from experiments involving changes in aspect to conjecture about the physiological underpinnings of seeing segregated units in general. He believes this is justified because his theory posits organization as a common explanatory factor for both aspects and other Gestalten. But, as we argued above, given a certain retinal stimulation, it may not even make sense to talk about seeing different Gestalten corresponding to physical objects. Thus, for these kinds of Gestalten, there seems little reason to consider a physical mechanism that was proposed in the first place specifically to explain *change*.

4.4 Conclusions

Wittgenstein and Köhler approach aspect seeing in distinctly different ways. While Wittgenstein wants to understand the phenomenon from

a conceptual point of view, Köhler is interested in giving a theoretical (and ultimately, physiological) explanation for it. By focusing on Köhler's theory of perception, and Wittgenstein's response to it, this Paper has therefore presented a kind of historical case study that shows how the scientific and the conceptual investigative paradigms can interact. In particular, Wittgenstein's remarks regarding the use of language expose weaknesses, and also an insight, in Köhler's account.

One of the shortcomings involves Köhler's notion of organization in itself. Köhler intends for this to be an explanatory concept. But upon reflection, we see that he cannot specify it other than as 'organization'. It therefore lacks explanatory content within the context of Köhler's project. This difficulty highlights a general need to reflect on the terms that are introduced into theories, in order to determine whether the terms really do their intended jobs. Conceptual reflection is especially pertinent when terms have an origin in everyday language and are then re-purposed for science. As we have seen with Köhler, just because a word has a certain use in some common contexts (such as a transitive use of 'organization' with respect to the structure of companies), this does not entail that the word automatically has the same use in other contexts (for example, with respect to psychological experiences).

A second shortcoming is that Köhler's theory, which invokes an organized visual field as a sensory fact, does not allow for the conceptual dissimilarity that exists in perception with respect to voluntariness. Although it is typically desirable for a psychological theory to account for as many related phenomena as possible, Köhler's case shows that it is also important not to overlook conceptual differences between the phenomena when developing such a theory. For Köhler seeks to explain *both* seeing segregated units that correspond to physical objects *and* seeing aspects of ambiguous figures like the pie by *one* underlying mechanism, namely, organization of the visual field (which, in turn, he says is a sensory fact like colour). In doing this, however, Köhler illicitly groups together cases that are conceptually distinct with respect to voluntariness. His attempt at giving a coherent treatment of seeing therefore fails on conceptual grounds before any data are even gathered.

Wittgenstein's conceptual studies of aspect seeing thus result in criticisms of Köhler's theory. But these criticisms pertain to specific features of the Gestalt psychology enterprise, and therefore they should not be taken as undermining (or as attempting to undermine) the overall value of the Gestalt tradition. For example, as we have discussed, Wittgenstein

explicitly acknowledges the validity and the worth of Köhler's impulse to talk of 'organization' of the visual field. If someone says 'When I look at the pie figure, my visual image is organized in a certain manner, and then the organization changes', we *do* understand what experience the speaker has had (even though the explanatory power Köhler wants is lacking). But more generally as well, Wittgenstein's thinking also exhibits an affinity to Köhler's programme with respect to the role of 'wholes' in perception. In Gestalt psychology, much importance is placed on wholes. This is witnessed, among other ways, by Köhler's acceptance of segregated entities (Gestalten), rather than mere colour patches, as basic in perception. A similar focus can be found in Wittgenstein's writings, such as when he comments, 'It might be like this: An eye can smile only *in a face*, but only in the entire figure can it—' (LW I, §860, emphasis original).[14] Here Wittgenstein's remark emphasizes that the entire face (or indeed the entire figure) that we see has a kind of primacy over the components, since a *smiling* eye (a part) requires the surrounding context of a face (the whole). Investigating this commonality between Gestalt psychology and Wittgenstein would take us outside the scope of this Paper. But in closing, we do want to stress that although Wittgenstein's philosophical remarks unmask conceptual weaknesses in Köhler's scientific proposals, the relation between Wittgenstein's work and Köhler's work is not characterised only by differences and by criticism.

Notes

1. There are also other cases of overlap between the examples discussed by Wittgenstein and by the Gestalt psychologists. For instance, Wittgenstein refers to Koffka's 'sail' in LPP, p. 332, he uses Koffka's 'mountains' figure in LPP, p. 337, he draws Köhler's 'cogs' in LPP, pp. 111 and 340, and he mentions Köhler's map of the Mediterranean in LPP, pp. 102, 109, 332 and RPP I, §1035 (cf. Koffka, 1935, pp. 129, 163; Köhler, 1947, pp. 152, 181). Here, and in the remainder of the Paper, we cite the more widely-available 1947 edition of Köhler's *Gestalt Psychology*. The 1929 version is not significantly different.
2. It is possible to argue for this claim, but doing so would take us beyond the scope of this Paper.
3. To put this differently: If a sighted speaker A looks at a face (when the face is nearby, in good lighting conditions and so on) and says, 'I don't see a face', then this would be a reason to ascribe some sort of impairment or disorder to A.
4. Putting it another way, as Peter Hacker (his emphasis) explains in the Prologue to the present volume, '[Philosophy] is a conceptual tribunal, whose role it is to tell people whether their questions, assumptions and conclusions make sense (not whether they are true or false), and, if they make no sense, to explain *why* they make no sense'.

5. It seems that William James (1842–1910), for instance, adhered to the Introspectionist picture that Köhler describes (cf. James, 1890, chapters XIX, XX).
6. See Budd (1989, pp. 83–6) and ter Hark (1990, pp. 175–8).
7. A paper by Cora Diamond provides an exception to the tendency for Wittgenstein's transitive/intransitive distinction to be overlooked. Diamond (2001) does not use the distinction to discuss experience or psychology, however, but rather Wittgenstein's remarks about the length of the standard metre bar. In a more recent paper, which came to our attention after we had written the present article, Avner Baz in fact discusses the transitive/intransitive distinction with respect to aspect seeing. But the specific concerns of Baz (2011) are different from ours.
8. In what follows, we use the terms 'visual image', 'visual impression' and 'visual field' interchangeably and assume that Wittgenstein does also (although in other contexts, he uses 'image' in the sense of what one imagines).
9. Consider also a parenthetical comment that Wittgenstein makes about inner objects and seeing aspects: '(The temptation to say "I see it like *this*", pointing to the same thing for "it" and "this")' (PPF, §214, emphasis original).
10. One might want to claim more specifically that the word 'organization' in the context of aspect seeing has what Wittgenstein calls 'secondary meaning' (cf. PPF, §§274–8; RPP I, §§125–6). That is: The concept involved when we speak of the organisation of our visual field is simply our ordinary concept of organisation. But the word 'organization' is used or applied differently with respect to aspect seeing than with respect to, for instance, the structure of companies. (For example, the word only has an intransitive use with respect to aspect seeing, whereas in the latter contexts it has a transitive use.) It is then our shared use of 'organization' with respect to aspect seeing that renders sentences like 'The organization of my visual field has changed' intelligible.
11. For example, Budd discusses being subject to the will with respect to both images (that is, what we have when we imagine) and aspects. In addition, he criticises Köhler on other grounds. Yet, he does not combine these two strands (cf. Budd, 1989, pp. 83–6, 94–5, 104–9). In a different vein, John Benjafield discusses some conceptual facets of voluntary action and correspondingly criticises certain empirical studies of 'reversals' of the Necker cube (2008, pp. 112–4). But Benjafield does not mention Wittgenstein's remarks about the voluntariness of aspect seeing or any possible applications to Köhler's work.
12. Interpreting the proviso 'at least to a certain extent' is beyond the scope of this Paper. But this added condition, which is also present in RPP I, §971, does not alter the argument we want to give regarding Köhler.
13. Distinguishing between conceptually different objects of sight is an important issue in Wittgenstein's work with aspect seeing. Consider, for instance, the differences between faces and likenesses that we have mentioned in our discussion of PPF, §111 in Section 4.1. For other arguments (not involving voluntariness) that criticise Köhler's failure to distinguish conceptually different objects of sight, see Cook (1994, pp. 142–4) and Schulte (1993, pp. 81–2).

14. It is noteworthy that Wittgenstein's remark in LW I, §860 in its original German uses *Gestalt* for 'figure'.

References

A. Baz (2011) 'Aspect Perception and Philosophical Difficulty' in O. Kuusela and M. McGinn (eds) *The Oxford Handbook of Wittgenstein* (Oxford: Oxford University Press), 697–713.

J. G. Benjafield (2008) 'Revisiting Wittgenstein on Köhler and Gestalt Psychology', *Journal of the History of the Behavioural Sciences*, 44(2), 99–118.

M. R. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Oxford: Blackwell Publishing).

M. Budd (1989) *Wittgenstein's Philosophy of Psychology* (London: Routledge).

J. W. Cook (1994) *Wittgenstein's Metaphysics* (Cambridge: Cambridge University Press).

C. Diamond (2001) 'How Long Is the Standard Meter in Paris?' In T. McCarthy and S. C. Stidd (eds) *Wittgenstein in America* (Oxford: Oxford University Press), 104–39.

M. R. M. ter Hark (1990) *Beyond the Inner and the Outer: Wittgenstein's Philosophy of Psychology* (Dordrecht: Kluwer Academic Publishers).

W. James (1890) *The Principles of Psychology*, Vol. 2 (New York: Henry Holt and Co.).

K. Koffka (1935) *Principles of Gestalt Psychology* (New York: Harcourt, Brace and World, Inc.).

W. Köhler (1929/1947) *Gestalt Psychology* (New York: Liveright Publishing Corporation).

—— (1940) *Dynamics in Psychology* (New York: Liveright Publishing Corporation).

R. Monk (1990) *Ludwig Wittgenstein: The Duty of Genius* (New York: The Free Press).

J. Schulte (1993) *Experience and Expression: Wittgenstein's Philosophy of Psychology* (Oxford: Oxford University Press).

5
Parallels in the Foundations of Mathematics and Psychology

Meredith Williams

The *Philosophical Investigations* ends, famously, with an indictment of scientific psychology, claiming that 'an investigation is possible in connexion with mathematics which is entirely analogous to our investigation of psychology' (PI II, p. xiv). Wittgenstein goes on to propose that the foundations of mathematics '...is just as little a *mathematical* investigation as the other is a psychological one.' It will *not* contain calculations, so it is not for example logistic. It might deserve the name of an investigation of the 'foundations of mathematics'. In this chapter, I want to take up the invitation to pursue the analogy, not only in terms of methodology, but also in terms of substantive similarities between the foundations of mathematics and the foundations of psychology. Mathematics and psychology are very different practices, but comparing foundational features is surprisingly illuminating. As Wittgenstein says earlier in the *Investigations*, 'the kind of certainty is the kind of language-game' (PI II, pp. xi, 224). So, to investigate the kinds of certainty belonging to these two language-games will reveal similarities and differences.

5.1 Platonist and Cartesian certainty

The certainty we have in mathematics and in psychology has led philosophers to look for special kinds of entities or realities lying behind the surface appearances: calculating in mathematics and wincing and crying in pain. Actual calculating, the writing of numerals on a page, is a temporally extended event subject to different kinds of failing. In adding, we may fail to carry a number over to the next column. We may take an 'X' for a '+'. We may make some random error. We may make some systematic error. Equally, crying and wincing might not be genuine. The behaviour might occur in the context of a play. It may be

deceitful. It may be an exaggerated behaviour. For all that, the certainty of mathematical propositions is not contingent (for that would only be psychological certainty). And the certainty secured by introspection is essential to the mental states in question. The certainty of mathematics and introspective awareness seems to reside in the very nature of mathematics and conscious experience. Further, this certainty is not undone because of the fallibility of actual calculations and the possibility of deceit with respect to our sensations, emotions and intentions. There is a problem, then, of showing how the propositions belonging to these two domains are certain while nonetheless being compatible with our fallibility in calculating and the fact that the certainty in ascribing psychological states to others is open to manipulation.

The traditional solution to the problem of certainty, and the solutions of Bertrand Russell (1912) and William James (1918), is to hypothesize the existence of metaphysically distinct realms. In the first instance, it is a metaphysical problem in accounting for the certainty of mathematics, not an epistemological one. The Platonist theory posits the existence of a realm of objects and numbers that exists independently of our calculations. Our calculations are the contingent means we have for constructing mathematical propositions. Thus, the conflict between contingent fallibility and absolute certainty is apparently resolved. Cartesian certainty, on the other hand, is epistemological; yet, behavioural expressions of pain or emotion or intention are only contingently connected to the mental states of a subject. Here the certainty of psychological knowledge must be made compatible with a contingent knowledge of the states of others. The solution is found in the idea that there is a domain of mental objects with special properties. These properties bring with them a principled distinction between first-person and third-person usage. First-person avowals are certain for the subject. The special properties of mental ideas ensure this certainty. Consequently, others have only a contingent grasp. Again, a metaphysical theory of special objects, ideas, is used to save a privileged place for certainty while acknowledging the problems raised by deception or failure to understand the mental states of others. Numbers and ideas, given their very special features, are utterly detachable from the contingent attempts to capture them in calculation or in the behavioural ties that express mental states. Only if they are so detachable can the certainty and necessity, which is essential to them both, be possible. But this solution comes with an immediate problem. The metaphysical theory of mathematics makes it impossible to understand how a mathematical or logical proposition could ever be applied.[1] The metaphysical theory

of mind makes it impossible to see how we could ever escape from the egocentric predicament.[2]

The reifying moves in the metaphysics of mathematics and mind cannot but fail, Wittgenstein argues, for they have mistaken norms for entities. The constraining consequences of norms are taken to be the hardness or luminescence of the special objects that inhabit the mathematical domain or the mental domain. Mathematical propositions and psychological propositions do not describe special objects within their respective domains, but express norms that are implicit in our mathematical and psychological language-games. Propositions like 'The number of primes is infinite', '$12 \times 12 = 144$', '1, 2, 3, and so on' or 'No one can share my sensations', 'Mental imaging is subject to the will', 'There is no such thing as reddish-green', on the metaphysical interpretation describe the nature of number or the nature of sensations and images. Wittgenstein argues that these propositions make explicit the norms that govern our judgements and inform our actions. In other words, for Wittgenstein, an investigation into the foundations of mathematics or psychology is an investigation into the normative constraints that structure these practices.

A grammatical, or philosophical, investigation is an investigation into the norms that are implicit in our language-games, norms that can only be realized within a community of practitioners and that are acquired in initiate learning. Grammatical statements make these norms explicit. They present norms. It is important to remember that these grammatical statements (like '$12 \times 12 = 144$' and 'Sensations cannot be shared') are normative because they *are* norms, not because they use rule-governed expressions. All statements are normative in this latter sense. Philosophy, in identifying grammatical propositions, that is, grammatical norms, is engaged in a descriptive enterprise. As norms, they express what is necessary and certain within the language they govern. Though Wittgenstein does endorse a strong distinction between grammatical propositions and empirical propositions, it is also clear that the line distinguishing them is blurred. Sometimes propositions that look like empirical statements in fact play the functional role of norms within its language-game. For example, in *On Certainty*, Wittgenstein takes the empirical proposition 'Napoleon existed', as a proposition held fast, that is, as a certainty within history (OC, §§163, 185). Similarly, the mathematical statement we have used as an example, '$12 \times 12 = 144$', is both a calculation at a particular time and an exemplar of mathematics. What matters to the normative status of any proposition is its role with in a language-game.

From a different direction, there has been a temptation to treat Wittgenstein's grammatical propositions as analytic truths, true in virtue of the meanings of the words. But this is a mistake. Wittgenstein has no notion of analytic truth. The distinction between analytic and synthetic statements presupposes a theory of meaning, which treats meanings as fully delineating, in virtue of intrinsic properties, the boundaries of use. In his attack on analyticity, Quine refers to this conception of meaning as the 'museum' theory of meanings (Quine, 1953, 1960). Meanings are labelled objects. Norms are not meanings in this sense; neither are they truths. They do not describe a reality either correctly or incorrectly. Rather, norms constrain the range of possible actions and judgements that can be made in the way that the rules of a game constrain action.[3]

Wittgenstein discusses the foundations of mathematics and psychology in parallel. He seeks to distinguish experimental or empirical propositions from timeless grammatical propositions (in mathematics) and psychological statements from logical (or conceptual) ones. He frequently uses the expressions 'mathematical' or 'geometrical' to characterize psychological propositions that are logical or grammatical, as in the following passages:

> 'There's no such thing as reddish green' is akin to the proposition that we use axioms in mathematics. (RPP I, §624)
>
> Suppose we explained this by saying the aspect comes about through different images and memories superimposed on the optical image. Naturally this explanation does interest me, not as an explanation but as a logical possibility, hence conceptually (mathematically). (RPP I, §1005)
>
> Under what conditions the language-game with the names of colours is physically impossible – i.e., properly speaking, not probable – does not interest us.
>
> Without chess-men one can't play chess – that is the impossibility which interest us. (RPP II, §199)
>
> And if the play of *expression* develops, then indeed I can say that a soul, something inner, is developing. But now the inner is no longer the cause of the expression (No more than mathematical thinking produces calculations, or is the impetus behind them. And this is a remark about concepts). (LW I, §947, emphasis original)

I chose these passages because they exemplify both the importance of distinguishing between the conceptual (logical) and the empirical and the analogy Wittgenstein discerns between mathematics and psychology.

The distinction Wittgenstein is drawing holds between propositions that state activities and propositions that are themselves norms or make norms explicit. Wittgenstein describes the proof for the infinity of prime numbers as providing a recipe for producing prime numbers.[4] Some logical propositions of psychology, like 'Imagination is subject to the will' or 'Sensations cannot be shared', express identity conditions. Neither the mathematical propositions nor the psychological ones are discoveries about a domain of special objects, numbers or sensations. It is not that one might discover a shared sensation as one might discover a black swan, or that the collection of primes really is finite or that mental imagery is produced by our sensory apparatus. On the contrary, the logical propositions are constitutive of their language-games in just the way the rules of chess are constitutive of the game.

Once we recognize that grammatical (or logical or conceptual) propositions are norms, then we can see that the necessary or internal connection between psychological expression and behaviour or between a mathematical function and a result is not to be explained metaphysically. Rather, the internal connection is that which is effected by a rule that is constitutive of a game. This is what Wittgenstein is stressing when he says:

> It looks like obscurantism to say that a calculation is not an experiment. And in the same way so does the statement that mathematics does not treat of signs, or that pain is not a form of behaviour. But only because people believe that one is asserting the existence of an intangible, i.e., a shadowy, object side by side with what we all can grasp. Whereas we are only pointing to different modes of employment of words. (RFM, §76)

Once we have accepted this distinction in the roles played by logical/conceptual propositions and empirical propositions, then we can see that there is no *metaphysical* difference in the kind of necessity that connects the outcome of a mathematical operation and the necessity expressed by the criteriological connection between sensation and behavioural expression. Getting that result and behaving that way (in those circumstances) are necessary in the way that chess-men are required for chess: 'Without chess-men one can't play chess – that is the impossibility which interests us.' Without behavioural expression, one can not ascribe sensation, not even in one's own case. This is shown by the Private Language Argument. Without getting 144, one can not multiply 12 by 12. To have the concept is to master these inferential and

logical connections that enable moves within the language-game. As Wittgenstein develops this in *RFM VI*, coming to grasp or master these concepts is coming to see that one must get such arithmetic results or recognize, in these circumstances, that one so acting is in pain. This necessity rules out certain possibilities and certain doubts. 'Just try, in a real case of fear or pain, to doubt it' (PI, §303).

5.2 Alternatives: pragmatism and naturalism

To bring out the originality of Wittgenstein's approach to these matters, we need to consider his view of the standard alternatives to the metaphysical accounts of Platonism or Cartesianism. The first strategy is pragmatism. The identifying claim of classical pragmatism treats norms as conventions or rules that we accept because it pays to do so. The second strategy is reductive naturalism, according to which the so-called norms of our language-games are reducible to natural facts of our physiology and ways of interacting with the natural environment. Wittgenstein is clear in his rejection of both of these familiar alternatives to metaphysics. Although we do sometimes think or count because it pays to do so, a pragmatist justification of our bedrock norms is mistaken.[5] Equally, the naturalistic reduction is mistaken, even though certain facts of nature and our own natural responsiveness are required for the language-games we sustain. Pragmatism, in looking for justification of our foundational 'norms', fails to understand the character of 'foundational' belief and action. Such beliefs and actions are decidedly not exercised for a reason. Reductive naturalism fails in a different way. It treats the empirical conditions for our language-games as exhaustive of those games.

Let us begin with Wittgenstein's critique of pragmatism. Wittgenstein is quite sensitive to the philosophical temptation to try to explain or justify the norms we have by appeal to pragmatic considerations: 'Why do we count? Has it proved practical? Do we have our concepts, for example, the psychological ones, because it has proved to be advantageous?' (RPP I, §951; Z, §700). Similarly, with respect to propositions that hold fast that Wittgenstein introduces in *On Certainty*. These propositions are held fast by the judgements and actions we take as a matter of course within our various language-games, not by our having found them advantageous. Yet, in his discussion of propositions that hold fast, Wittgenstein is aware of the danger of being misinterpreted as a kind of pragmatist: 'So I am trying to say something that sounds like pragmatism. Here I am being thwarted by a kind of *Weltanschauung*' (OC, §422). That 'world picture' is one in which we human beings choose or could

choose our conventions and norms (as we chose the standard meter stick) in accordance with sound pragmatist principles. What Wittgenstein is struggling against is this optimistic picture of human control of the language-games we play. He counters the pragmatist ambition when he asserts that '[l]anguage did not emerge from some kind of ratiocination' (OC, §475). Choosing, justifying, weighing the consequences of our choices and actions are all the actions of a more sophisticated sort that are possible only against the background of normatively structured practices that are not a matter of choice or the result of justification. Pragmatism's intellectualism is misplaced according to Wittgenstein, and more deeply it is founded on a misunderstanding of the background to our language-games.

Wittgenstein holds that any language-game has a background that informs and constrains the particular moves within the language. There are three dimensions of the background: mastery of techniques, the context within which a particular use of language occurs, and propositions held fast. We engage in our ordinary uses of language in ways that are blind to these three dimensions. Reflection, however, may bring these features to light. Wittgensteinian philosophy aims at describing some of these features. This reflection is not like that of the pragmatist, who thinks that description of the background opens the background to revision. This is how pragmatism goes wrong. The background is one that requires blind obedience, obedience that is not a sign, however, of our irrationality or gullibility: 'And that something stands fast for me is not grounded in my stupidity or credulity' (OC, §235).

The place to look for a proper understanding of the normative background is the initiate learning situation. The false ambitions of the pragmatist are most clearly revealed by comparing the naive learner to the fully linguistically competent. Initiate learning and trust form the milieu in which linguistic techniques are acquired, and bedrock is held fast. One distinctive feature of this kind of learning is that the linguistic mastery that is acquired is rooted in an absolute certainty that is the complement of the blind trust that the novice places in the master of the practice. Here is where certainty finds its home and never leaves the novice. Certainty is not achieved through a special kind of introspection nor can it be replaced by causal necessity. Certainty grows out of the absence of alternative ways of behaving and the blind trust the novice holds for the adult. This is not a form of ratiocination. The young child does not believe anything but, at best, has proto-beliefs that are identified with the beliefs of the adult who provides the cognitive background for the utterances and behaviours of the child. That is, the adult treats

the child *as-if* he were a cognitive agent.[6] Becoming a believer or one who expects or intends is a gradual process that mirrors the gradual process of coming to understand. The adult provides support for the child's utterances before the child is capable of making genuine judgements or acting in a norm-governed way.

Wittgenstein takes the special relation between child and adult, novice and master, to be a logical point about the structure of belief and the initiate learning situation. It is not just an empirical hypothesis that the novice needs a master. It is required, given the structure of the background without which there can be no language. If the child were a hypothesis formation 'machine', he could not learn language. He would have no way of providing the background to the language-games he engages in. Pragmatism, as a philosophy, fails to grasp these features of the background.

The reductive naturalist also fails to understand the background though in a different way. The naturalist seeks to show that the background is a set of causal properties that the agent has to his own physiology and to the natural environment. The naturalist denies a place for norms and normativity in the ways in which we participate in various language-games. There are two ways to challenge this picture. The child becomes a participant in the language-game he acquires. This is a matter of becoming subject to norms: To acquire mastery of skills necessary to play a game, to recognize the relevant from what is irrelevant while playing the game, and to thereby hold in place deeply entrenched grammatical propositions. The novice comes to understand the game. What the naturalist is correct about is that initiate training requires certain natural reactions on the part of the child.

The second challenge comes from another feature of normativity. The mark of engaging in a normative practice, whether a formal game or an inferentialist game, is the possibility of mistake or error. Failure to get the correct result for a multiplication problem can only be interpreted as either a failure to understand the arithmetic function multiplication or as a mistake made in the process of multiplying. These two possible ways of understanding mistakes presuppose that the agent is in normative space. Mistakes do not occur in the causal nexus. In a similar way, failure to behave in the ways appropriate to expecting someone to tea is a failure to expect anyone. Such normatively structured enterprises do not, for the most part, originate in rules as formulae or in exemplary objects. They originate in socially constrained behaviour, the learning of which (the acquisition of which) is a function of our shared natural responses to the environment and to each other within a context of

normatively informed practices. These ways of acting come to implicate background pictures that are held fast by the actions themselves. This is a matter of being the sort of creatures we are, of sharing a human form of life.

Both pragmatism and naturalism misunderstand the special character of the background of norms that inform every language-game. There is a further methodological commitment shared by pragmatism and naturalism. This is the view that our doings as language-users can be seen from outside, as it were. It is from that standpoint that pragmatism and naturalism seek to make their claims. Wittgenstein seems to make a similar presumption. He acknowledges that if one were to '... imagine certain very general facts of nature to be different from what we are used to, and the formation of concepts different from the usual ones will become intelligible to him' (PI II, p. xii). Wittgenstein grants that it is contingent that we have the language-games we do. There is no absolute necessity for any of them. Whatever necessity our language has is internal to the language-games themselves.

5.3 The analogies between mathematics and psychology

The analogies between mathematics and psychology concern, of course, their foundations. The surface structures of these language-games are manifestly quite distinct. The foundational questions concern the kind and source of the necessity crucial to each language-game, a necessity captured by logical or grammatical propositions rather than empirical ones. That is, this necessity is not that of causal connection or the instantiation of law or special metaphysical powers of mind and number. Wittgenstein is looking for what he calls a 'radical explanation' (RFM) for mathematical necessity and, by extension, the necessity that grounds our psychological language-games. The kind of necessity that is constitutive of the foundations of both mathematics and psychology is of the *same kind*. The source of necessity is explicated, in part, in terms of the social dimension of our language-games: in terms, that is, of agreement in bedrock judgements, whether of mathematics, psychological attribution or any other language-game.

Comparing psychological propositions with mathematical propositions sheds light on the nature of necessity found in each language-game. The mathematic proposition '$12 \times 12 = 144$' is such that getting '144' as the result of multiplying is necessary. One cannot have been multiplying correctly if one got a different number for the product. This proposition is an exemplar of mathematical necessity. Wittgenstein

makes two important points in connection with *mathematical* necessity: (1) bedrock mathematical propositions are normative, and that is why they are necessary; and (2) their necessity is deontic, not metaphysical. It is a matter of how the agent must go on. This is Wittgenstein's replacement explanation of metaphysical necessity. Using the mathematical exemplar, Wittgenstein applies to psychological judgements the same two points. The psychological judgement 'He is groaning in pain' also expresses, as the mathematical judgement did, necessary inference. He takes the groaning to be criterial for being in pain as an empirically based inference, Wittgenstein treats it as a necessary inference, that is, he takes the former to be *criterial* for the latter. The necessity that obtains here is no different from the kind of necessity that obtains between an arithmetic function and its product.

But the analogy runs in both directions. Mathematics illuminates the necessity of criterial judgements: the kind of necessity that is neither metaphysical nor causal. And criterial psychological judgements illuminate the source of mathematical necessity: in the bedrock techniques that are acquired in initiate learning, where 'initiate learning' is per force a training for those without linguistic mastery. This could be the child just learning her colour words or the pupil just learning to count. That '5' must follow '4' if one is adding by singletons, that (typical American) fire engines are paradigms for teaching or explaining the meaning of the word 'red', that the connection between groaning and pain is criterial: all express necessity. The kind of necessity and the source of necessity are displayed in the analogies between mathematics and psychology.

The mathematical proposition '$12 \times 12 = 144$' and the psychological proposition 'He is in pain', indicating a man groaning, are normative. Yet, they are not explicit norms. They do not say what we ought or must do within the language. But their *role* in our language-games is normative. The proposition is not up for revision. It is used as a standard for understanding basic arithmetic. Someone who doubts that $2+2=4$ has not learned his or her numbers. Propositions with a normative role are also found in our psychological language-games. This is an important point of analogy with mathematics and psychology. In psychology, such propositions with such a normative function can have the appearance of empirical propositions. Though they look like empirical propositions, they have their normative role in just the way that '$12 \times 12 = 144$' does for mathematics.

Wittgenstein thinks that the learning situation is especially fruitful for examining the structure or features of a normative practice. The learning

situation is one in which the novice acquires bedrock techniques that are the background for primitive judgements of similarity or sameness:

> 'How am I able to obey a rule?' – if this is not a question about causes, then it is about the justification for my following the rule in the way I do.
>
> If I have exhausted the justification I have reached bedrock, and my spade is turned. Then I am inclined to say: 'This is simply what I do'. (PI, §217)

In the learning situation, one must begin with bedrock. The novice or pupil lacks the cognitive resources to be open to that which is subject to justification. Primitive bedrock judgements are the novice's reaction. The structure of the learning situation is that of the relation between the novice and the master of language. The child makes the bedrock 'judgements' which are supported cognitively by the competence of the adult.

The technique of counting is one such basic technique in arithmetic. Primitive recognition of colours provides basic ways of measuring the world. Responding to a cry of pain is another. These are all natural responses to training within these three language-games. The special character of the learning situation shows how the causal factor is related to the normative. To the child who first learns these, bedrock judgements are akin to causation, but to the adult who trains the child, they are skills essential to their respective language-games. They provide bedrock judgements of similarity and sameness.

Initiate learning creates the background for the individual without which there are no moves within any language-game. Belief formation in initiate learning is of a different nature from that which occurs in a process of hypothesis formation and testing. The critical rational attitude that is part of the latter is necessarily missing from the former. The novice lacks the background skills or beliefs to form hypotheses, evaluate evidence or assign probabilities. The novice acts and judges on trust, with 'light dawn[ing] gradually over the whole' (OC, §141). The important feature of novice belief is that in accepting these judgements on trust, the novice does not select them from among several possibilities. Indeed, the novice lacks the resources to consider alternative ways of judging. In an important sense, the necessity of believing what one is taught is no more an option for the novice than is flying rather than falling when one trips over a stone. There are no alternative hypotheses or systems of belief available to the novice, other than the one on

offer from his teachers and community. As Wittgenstein emphasizes, this is not a failure of intelligence or curiosity or diligence on the part of the novice. Such alternatives can only become available against a background system (or systems) of judgement and action. This is not a psychological defect or handicap, but a logical feature of language-games. This background is one that the child acquires slowly and in interaction with those who already have linguistic competence.

To participate in a language-game is to act and judge against a background of techniques, certainties and surroundings. When that background does not exist for an individual, when the individual lacks the techniques for recognizing the obvious, the background certainties and/or the appropriate surroundings, then the individual cannot select or choose from an array of possible actions. The world does not have a logical form in which all possibilities subsist, independently of the particular practices that contingently exist. Wittgenstein's insight is that the form of believing and acting as a participant in a language-game involves both the contingency of the language-game, and so the possibility of other games, and the necessity of the believing, such that alternatives are not even available. Our being learners initiated into community practices reconciles these two features. The normative judgements and actions we acquire in initiate learning may as well be causally necessary. To play the game is to come to accept certain moves as the way things must be, as the way in which one acts as a matter of course. Forming this second nature is a matter of believing without grounds and without alternatives.

This is a logical point, not an empirical discovery or hypothesis. Here are two passages in which Wittgenstein is concerned with just this issue:

> Am I doing child psychology? – I am making a connexion between the concept of teaching and the concept of meaning. (Z, §412; also see RPP II, §337)
>
> Clearly the words 'Now I am seeing this as an apex – now that' won't mean anything to a learner who has only just met the concepts apex, base, and so on. But I did not mean this as an empirical proposition. (RPP II, §483; also see PI II, p. 208)

The way in which Wittgenstein relates the necessity of our language-games to the learning situation and the way in which he views grammar as grounded in our normative ways of acting are constitutive of all language-games, but we see this quite clearly in the case of

mathematics. To establish an analogy between the two is show how (grammatical) necessity infuses our psychological language-games as it does mathematics.

In the passage from *Zettel* just quoted, Wittgenstein raises the question 'Am I doing child psychology?' to which he replies, 'I am making a connexion between the concept of teaching and the concept of meaning.' When Wittgenstein says that there is a connection between the concepts, he does not mean a contingent empirically established relation. He means a conceptual or logical or grammatical connection. So, how does he argue for such a connection in mathematics and in our psychological language-games?

The 'learning' language-games provide a simple situation in which the child learns how to become an actor in the relevant language-games. These are natural thought experiments in which the child's initiation into a practice provides insight into the meaning of the words he uses. The background dimension that is highlighted in the learning mathematics is the technique for manipulating numbers in specific ways. The technique is a matter for knowing *how*; it is not a matter for knowing *that*. Mastery of a technique is shown in the context of the simple learning game. That is the best account that we can give of mathematical techniques. Techniques are undeniably essential to mathematics. Learning a technique is tied to the meaning of mathematics. Also implicated are the context created within the learning situation and the background certainties that are held fast by the moves that are sanctioned within the game.

When we turn to psychological language-games and ask of them how the learning situation relates to the meaning of the psychological words in play, we will look to a different dimension of the background to explain that connection. The context, *pace* mathematics, is highlighted for our acquisition of psychological terms. The learning situation in which the child acquires the use of psychological terms is a matter of serendipity. The child falls and hurts herself. The child is tickled and laughs. These are the sorts of situations in which the child can learn the words for pain and tickle for herself and for others. These, too, involve certain natural reactions on the part of a child. Crying, wincing, looking for help are all natural reactions to being hurt. Learning to use the words 'It hurts' is to acquire new pain-behaviour (PI, §244). Teaching in this context involves the child's being made aware of the vulnerabilities of her own body. As techniques are required for the meaningfulness of mathematics, so the behavioural and situational contexts are required for our psychological terms. Something like technique may be necessary to account for

certain expertise in judging behaviour from a psychological perspective. Equally, background certainties are inevitable as is the case with any language-game. But context is the background feature Wittgenstein is especially concerned with. Learning the concepts of pain, tickles and the like is acquiring the behavioural context in which pain or tickles occur. The child's primitive psychological learning requires behaviour which is expressive of the experience. Thus, the criterial status of the behaviour. Even in adult life, when deception is possible, that criterial status of behaviour in relation to a mental state is not undermined. That status, however, is defeasible and thus reveals itself to be different from that of technique. Even though criteria for a psychological state are defeasible, they nonetheless remain necessary.

The relevance of learning to meaning and the nature of necessity are two important analogies between the foundations of mathematics and psychology. There is a third:

> Meaning it is not a process which accompanies a word. For no process has the consequences of meaning.
>
> (Similarly, I think, it could be said: a calculation is not an experiment, for no experiment could have the peculiar consequences of a multiplication). (PI II, p. 218)

Meaning is not a process that accompanies a sentence or the writing of a calculation, because 'no process has the consequences of meaning.' This is certainly true of mathematics and psychology. Performances in both language-games create consequences that go beyond the formulae and statements used. A process as an event cannot have the consequences of meaning. Constructing a proof that the number of primes is infinite has consequences that go well beyond the process of constructing a proof at a particular time. More simply, counting a segment of the natural number sequence equally has consequences that go beyond the process of enumerating some segment. Psychological propositions also have consequences that go beyond the statement made, though these differ from that of mathematics. 'I'm expecting Mr NN to tea' has consequences of meaning going beyond linguistic process of saying, 'I'm expecting Mr NN to tea.' The consequences of a process cannot be the same as the consequences of the meaning of a statement. The consequences of the former require locating the event in the causal nexus, while the consequences of meaning require locating the meaningful sentence in the inferential space of action, reason and belief.

5.4 The disanalogies between mathematics and psychology

The disanalogies that Wittgenstein emphasizes pertain to the kind of certainty realized in each language-game and the kind of learning required for the mastery of each game. These two are interconnected. The certainty that is a logical condition of any game is believing on trust, that is, blindly obeying, and that is one of the distinctive features of initiate learning. Here is the first disanalogy:

> Am I less certain that this man is in pain than that twice two is four? – Does this shew the former to be mathematical certainty? – 'Mathematical certainty' is not a psychological concept.
>
> The kind of certainty is the kind of language-game. (PI II, p. 224)
>
> 'But you can't recognize pain with certainty just from externals.' – The only way of recognizing it is by externals, and the uncertainty is constitutional. It is not a shortcoming.
>
> It resides in our concept that this uncertainty exists, in our instrument. Whether this concept is practical or impractical is really not the question. (RPP II, §657)

These two passages have to do with the kind of certainty involved in each language-game. The difference is sharply put: '"Mathematical certainty" is not a psychological concept' and 'the uncertainty [in recognizing feelings in others] is constitutional ... it resides in our concept.' The certainty of mathematics is a logical feature, one revealed in the philosophical examination of the foundation of mathematics. Psychological concepts, on the other hand, have a certainty which is marked by a constitutional uncertainty with respect to the psychological states of others. That is the kind of certainty that psychological language-games have.

There are three dimensions along which the certainty of a language-game can be assessed: the degree of certainty, the kind of error or mistake, and the scope of agreement among practitioners of the game. The degree of certainty that mathematics requires is absolute. As indicated above, this is not a psychological certainty, say, a strong feeling of confidence, but a logical requirement concerning what results must be obtained in doing mathematics. Secondly, the kind of certainty has implications for the place of error or mistake in mathematics. Failure to get the proper results in mathematics is not just a different perspective, say, on the formulae; it is a failure of understanding. There is no

constitutional uncertainty constitutive of mathematics. These first two dimensions make the third inevitable. Agreement among those doing math must be universal:

> This consideration [about agreement] must ... apply to mathematics too. If there were not complete agreement, then neither would human beings be learning the technique which we learn. It would be more or less different from ours up to the point of unrecognizability. (PI II, p. 226)

These three features lay out the certainty that belongs to mathematics. The certainty of our psychological language-games is quite different.

The notable difference between our mathematical concepts and our psychological ones are that the latter are tied to the sincerity with which the subject speaks and acts. Being sincere in solving mathematical equations is irrelevant to the logic of mathematics. What matters is getting what anyone would get in solving the problem. Mastery of mathematical techniques is necessary. It is this essential link to sincerity that makes uncertainty constitutive of our psychological concepts. To pretend to feel something is analogous to making a mistake in mathematics. Both presuppose that much stands fast, and so not open to doubt, in order to make logical room for the possibility of either pretence or mistake. To make a mistake in one's calculations requires that one have mastery within mathematics. Similarly, Wittgenstein holds, I cannot pretend to suffer unless I have mastered an array of psychological concepts, a mastery that requires recognizing the essential connection between behaviours of certain sorts and psychological states of certain sorts. It is precisely by exploiting that understanding that pretence is possible at all.[7] Those who have mastered psychological concepts, have thereby the means for deceiving others. Logical room for deception is part and parcel of the mastery of psychological concepts. The first two dimensions of the certainty of psychological language-games are interrelated. Uncertainty in psychological language-games is constitutive of those games. The range of mistakes one could make about others is tied to this feature of these games.

The certainty of mathematics and the constitutive uncertainty of psychological concepts are mirrored in the degree of agreement required of participants in each of the games. Despite room for considerable disagreement over psychological attribution, the constitutional uncertainty presupposes and requires a certainty at bedrock level that mirrors mathematical certainty. Bedrock certainty, that is, the certainty that

reflects the essential (or criterial) connection between behaviour and feeling is part and parcel of the language-game itself. The disagreements are not over whether groaning is expressive of suffering, but whether the subject is sincere or not. The disagreements cannot be understood except against the background of the certainty of criterial connections. With the judgement of sincerity comes certainty in the attribution of a psychological state. Within mathematics, a failure to answer '5' to the question 'what is 2+3?' is seen as a mark of not understanding mathematics, since it is a failure to grasp a paradigmatic judgement. Likewise, with respect to our psychological attributions, '"You're all at sea!" – we say this when someone doubts what we recognize as clearly genuine' (PI II, p. 227). This kind of response, Wittgenstein is saying, is indicative of the violation of a norm for the practice itself. A great difference remains, of course, between these two practices. In mathematics we can provide a proof for a mathematical proposition, but in the realm of human psychology, 'we cannot prove anything' but must rely upon 'imponderable evidence' such as 'subtleties of glance, of gesture, of tone' (PI II, p. 228).

These differences of the kinds of certainty are reflected in how we become participants in the two games. In the following passage, Wittgenstein indicates the relation between the complete agreement we have in mathematics and judgements of colour (this could equally well have judgements of pain) and learning techniques:

> Does it make sense to say that people generally agree in their judgments of colour? What would it be like for them not? ...
>
> How would they learn to use these words? And is the language game which they learn still such as we call the use of 'names of colour'? ...
>
> This consideration must, however, apply to mathematics too. If there were not complete agreement, then neither would human beings be learning the technique which we learn. It would be more or less different from ours up to the point of unrecognizability. (PI, II, p. 226)

How we are taught mathematics and psychological concepts is quite different, reflecting the different kinds of certainty and need for agreement involved in each:

> 'We all learn the same multiplication table.' This might, no doubt, be a remark about the teaching of arithmetic in our schools, – but also an observation about the concept of the multiplication table. ...

> There is in general no such agreement over the question whether an expression of feeling is genuine or not. ...
>
> Can one learn this knowledge [of the genuineness of expressions of feeling]? Yes; some can. Not, however, by taking a course in it, but through *'experience'*. – Can someone else be a man's teacher in this? Certainly. From time to time he gives him the right *tip*. – This is what 'learning' and 'teaching' are like here. – What one acquires is not a technique; one learns correct judgments. There are also rules, but they do not form a system, and only experienced people can apply them right. Unlike calculating rules. (PI II, p. 227, emphasis original)

These differences emerge with the kind of training that novices receive into the language-games. In mathematics, one learns techniques for using the various mathematical operators, techniques that do not permit different individuals to reach different conclusions. Indispensable to this form of teaching is the fact that we do respond in a common manner to this teaching. Initiate learning is the acquisition of normative beliefs that create the logical space of numbers. There are psychological language-games that are similar to mathematics in this regard. The primitive language-games in which we teach young children simple psychological concepts tie the feeling or desire to simple behaviours. The strong criterial connection between sensation or primitive desire or belief and contextualized behaviour remains a bedrock of certainty against which the sophisticated psychology of adults can be measured. This is reflected then in how we are taught such a complex game. We are taught in terms of good judgement, not by way of techniques that ensure an outcome. We have no proofs for the psychological states of others (or ourselves), only the imponderable evidence missed by those who are blind to such nuances of feeling. Those who come to form expert judgement do so only as a result of experience and 'tips' from others adept in such matters.[8]

5.5 Conclusion

Wittgenstein's philosophical investigations into the foundations of mathematics and the foundations of psychology are inquiries into the deep background of language-games. Wittgenstein's task is to an extent a Kantian one. He is seeking to uncover the conditions that make possible commonly shared human forms of life, most especially, our linguistic relation to ordinary objects, mathematics, and psychology. In each of these discussions, Wittgenstein seeks to understand how both

the builder and the logician are involved. The primitive calls of the simple builders are as important to our form of life as the sophisticated constructions of logicians. The challenge is to understand how both calls and abstract signing are implicated, for all of us, in our use of language. To try to understand language use in terms of one only inevitably results in distortion.

Notes

1. Of course, Wittgenstein was already acutely aware of this difficulty in the *Tractatus*. Indeed, much of the *Tractatus* is concerned with removing just this problem by rejecting the metaphysical thesis.
2. Wittgenstein was already grappling with both of these problems in the *Tractatus*, solving the first, as he thought then, with the disappearance theory of logical constants and the second with the removal of the subject from the world.
3. If one were to press continuity between the *Tractatus* and the *Investigations*, one might say that norms replace the Tractarian idea of logical form. Where there is a single universal logical form that fixes the range of possible states of affairs, the Investigation invokes a multiplicity of language games, each of which constrains the range of possible moves within the game. Even this is more stringent that Wittgenstein actually demands.
4. For a fuller discussion, see Williams (2010, ch. 7).
5. Contemporary pragmatism has evolved in ways that break with some central ideas of classical pragmatism, especially the conception of truth. Many contemporary pragmatists reject the idea that truth is what it is useful to believe in favour of a deflationary theory of truth. See Brandom (2011).
6. See Clark (2009). Though disagreements concerning first language acquisition persist, Clark's treatment is thorough and does not conflict with the features of initiate learning that Wittgenstein cites. The situation in which a child can learn language is not one of merely overhearing adults speak. Rather the situation is structured. A social setting is required in which the adult engages in 'child directed speech' (p. 22). Typically, this can involve what philosophers have called 'triangulation' (p. 27). This is a situation in which the adult, the child, and an object are triangulated, so that the object is the focus of both the adult and the child. This requires an I-thou relation to obtain between child and adult. All of these describe Wittgenstein's initiate learning situation. One final point: Infants engage in indicating and requesting gestures that are best thought of as 'protospeech acts, the forerunners of the speech acts with those functions' (p. 88).
7. Wittgenstein makes this point indirectly when he asks, 'Are we perhaps over-hasty in our assumption that the smile of an unweaned infant is not a pretence?' (PI, §249). He follows this up by asking, 'Why can't a dog simulate pain? Is he too honest?' (PI, §250).
8. For a more sustained and subtle treatment of the issues, see Schulte (1993), esp. ch. 4 'Expression', and Johnston (1993), esp. ch. 4 'The Musicality of Language'.

References

R. Brandom (2011) *Perspectives in Pragmatism: Classic, Recent, and Contemporary* (Cambridge, MA: Harvard University Press).
E. V. Clark (2009) *First Language Acquisition*, 2nd edn (Cambridge: Cambridge University Press).
W. James (1918) *The Principles of Psychology*, Vol. I (New York: Dover Publications).
P. Johnston (1993) *Wittgenstein Rethinking the Inner* (New York: Routledge).
W. V. Quine (1953) 'Two Dogmas of Empiricism', *The Philosophical Review*, 60, 24–43.
—— (1960) *Word & Object* (Boston: MIT Press).
B. Russell (1912) *The Problems of Philosophy* (Oxford: Oxford University Press).
J. Schulte (1993) *Experience and Expression* (Oxford: Clarendon Press).
M. Williams (2010) *Blind Obedience* (London: Routledge).

6
Animal Minds: Philosophical and Scientific Aspects

Hans-Johann Glock

This essay discusses the relation between philosophical and scientific aspects of the topic of animal mentality. It defends the method of conceptual analysis both in general and with respect to the topic of animal minds (Sections 6.1–6.4). But it also argues for a type of conceptual analysis that is non-reductive and impure (Sections 6.5–6.6). This approach distinguishes the conceptual issues of philosophy from the factual issues of science, while being sensitive to the way in which these interact in specific questions, arguments, theories and research programmes. Philosophy is distinct from science, yet the two cannot proceed in isolation with respect to topics like that of animal minds, which pose both scientific and philosophical problems. A more specific reason for favouring impure conceptual analysis in the philosophy of psychology, but especially with respect to animal minds, is the importance of methodological issues which are neither straightforwardly conceptual nor straightforwardly factual (Sections 6.7–6.8). Section 6.9 dwells on the connection between the proper analysis of mental concepts and our practice of applying many of them to animals.

6.1 Animal minds, philosophy, and conceptual analysis

The starting point of this inquiry is the question of animal minds:

> Do some animals have minds/mental properties/mental powers?
>
> Or, to put the problem in a more specific manner:
>
> What mental properties, if any, are possessed by what species of animals?

Call this the distribution question (see Allen, 2010). Just as the question of animal minds and the distribution problem concern the mind

or mental properties in general, there are also analogous questions concerning specific mental properties. The answer to such questions depends on two factors. On the one hand, it depends on contingent facts about animals to be established by empirical science, whether through observations in the field or experiments in the laboratory. On the other hand, it depends on what one makes of heavily contested concepts like that of a mind, of thought, rationality, consciousness, perception, sensation or behaviour. Let us assume that all the empirical facts about animals – their behaviour, neurophysiology and evolutionary origins – have been established or that they can at least be taken for granted for the sake of argument. What these facts imply for the possession of mental properties would still depend on what these properties amount to. Conversely, if all the pertinent concepts/properties have been determined, which animals actually satisfy the concept/possess the properties will depend on contingent facts.

This involvement of both factual and conceptual factors in the question of animal minds is truistic and applies more generally to empirical questions employing concepts that are ambiguous, vague or contested. To vindicate the method of conceptual analysis, however, requires more by way of support than semantic truisms. First, it is not just that both factual and conceptual factors impinge on the proper answer to empirical questions, but also that these two factors can be held apart. Secondly, whereas empirical science tackles factual issues, conceptual issues constitute the (or at least a) central purview of philosophy.

For reasons of space, I cannot defend the second claim here (see Glock, 2009b, pp. 340–1). In this essay, I shall assume that the main philosophical task regarding animal minds and psychology more generally consists in clarifying mental concepts, notably by investigating their conditions of application, the conditions which something must fulfil to satisfy these concepts. Empirical science, by contrast, determines whether or not these concepts do, in fact, apply to animals of a certain species; it also provides causal explanations of how they come to satisfy these conditions of application. In addition to the application and the elucidation of concepts, there is conceptual construction or concept-formation, the devising of novel conceptual structures. This activity is one of the hallmarks of mathematics, which provides novel formal tools for describing and explaining empirical phenomena. But it features in all forms of discourse, since it occurs whenever new ways of classifying or explaining phenomena, of thinking about or making sense of them are introduced. Philosophy is no exception; the concepts of analyticity and

a priority, for instance, are philosophical innovations. Yet, the formation of entirely new concepts for the description, explanation and prediction of (physical) reality is the task of the empirical disciplines that try to make sense of the pertinent data.

6.2 Conceptual analysis and definition

Allotting to philosophy the task of clarifying concepts is the trademark of a particular current within analytic philosophy, namely conceptual analysis (see Glock, 2008). Furthermore, conceptual elucidation is an activity for which contemporary analytic philosophers seem particularly well equipped. The idea that philosophy seeks to analyse or define concepts, rather than to decide what they actually apply to on the basis of experience has a venerable pedigree. Ever since Socrates, philosophers have been concerned with 'What is X?' or 'What are Xs?' questions, for example 'What is justice?', 'What is knowledge?', 'What is truth?'. In response to these questions, they have traditionally sought analytic definitions of X(s). Such definitions specify conditions or features which are individually necessary and jointly sufficient for being X. Furthermore, these features should not just, in fact, be possessed by all and only things that are X; rather, it should be necessary that all and only things that are X possess them. Only things possessing all of the defining features can be X, and anything possessing them all is, ipso facto, X.

Analytic definitions can be understood either as nominal definitions, which specify the linguistic meaning of words, or as real definitions, which identify the nature or essence of the things denoted by them, something independent of the way we think and speak. Both traditional and contemporary metaphysicians have sought real definitions capturing the mind-independent nature or essence of things. They purport to discover substantive truths about reality that are more general and fundamental than those of science and to ascertain the nature or essence of things, yet without relying on experience. As Kant pointed out, this ambition is puzzling. How can we achieve 'synthetic' insights about reality independently of experience? To this day, the Kantian challenge awaits a compelling response (Glock, 2009b). Within the analytic tradition, the idea of *de re* essences has been rejected by figures as diverse as Wittgenstein, the logical positivists, ordinary language philosophers, Popper and Quine. In their wake, many twentieth-century philosophers settled for nominal or *de dicto* definitions, definitions or explanations which specify the meaning(s) of 'X'.

Alas, at present, such a procedure will be greeted by the indignant complaint that philosophy ought to be interested not in 'mere words' but rather in the nature of the things they denote. Any sane proponent of conceptual analysis will recognize the difference between words and concepts on the one hand, and the things they denote or apply to on the other (Hanfling, 2000, p. 17; White, 1987, ch. 2). Nonetheless, the question 'What is X?' often concerns the meaning of words. Admittedly, questions of this form can be requests for empirical information about contingent features of X(s). But 'What is X?' questions as posed by philosophers are not directed at contingent features of X, features which instances of the concept may or may not possess and which need to be established empirically by looking at these instances. They are directed instead at the essence or nature of X, at what makes something an X in the first place. And that kind of question is properly answered by an explanation of what 'X' means. For such an explanation will specify what counts as X, or what it is to be an X, independently of features that X may or may not possess and which need to be established empirically. Similarly, for questions of the form 'What makes something (an) X?'. These can be requests for a causal explanation that specifies how come that certain objects are X. But they can also be requests for a semantic explanation that specifies what constitutes being an X, that is conditions by virtue of which something qualifies as (an) X in the first place. There is no significant difference between saying – in what Carnap called the 'material mode' – that a drake is a male duck and saying – in the 'formal mode' that 'drake' means 'male duck'.

Still, a genuine gap between nature and meaning looms if one adumbrates the powerful revival of essentialism through Kripke (1980) and Putnam (1975, ch. 12). According to their 'realist semantics', the reference of natural kind terms like 'water' or 'tiger' is not determined by the criteria for their application that are specified in nominal definitions – the phenomenal features by which laypeople distinguish things as belonging to those kinds, such as the way something looks or tastes. Rather, it is given by a paradigmatic exemplar and an appropriate 'sameness relation' that all members of the kind must bear to this exemplar. 'Water', for instance, refers to all stuff which is relevantly similar to a paradigmatic sample, that is any substance which has the same microstructure as that paradigm. Accordingly, natural kinds do not just possess a 'nominal' but also a 'real essence', in Locke's terminology (1975, III.3), in this case to consist of H_2O.

Emboldened by realist semantics, many contemporaries proclaim that they can get *de re* essences into the crosshairs of their intellectual

periscopes. Yet, all the while they tacitly rely on their understanding of philosophically-contested expressions, both in rejecting others' accounts of the pertinent phenomena and in proposing their own alternatives. Although realist semantics seeks necessities which concern reality rather than our conceptual scheme, they identify these through the workings of language, notably the alleged fact that natural kind terms function as 'rigid designators' that refer to the same phenomena in all possible scenarios.

Whether the realist account actually fits our use even of natural kind terms like 'water', for which there are concrete paradigms that can be investigated by science, is doubtful nonetheless (Hacker, 1996, ch. 7; Hanfling, 2000, ch. 12; Glock, 2003, pp. 95–101). Such doubts are exacerbated when it comes to mental terms. For there is no thing, organism or stuff called, for example, (a) pain, consciousness or reasoning that one could point to as a paradigm and subject to scientific scrutiny. What one can point to are only paradigmatic behavioural expressions of these phenomena, that is to cases in which the criteria specified by a nominal definition are fulfilled. This leaves open the possibility that mental phenomena have underlying scientific essences, namely the micro-structural neurophysiological causes of behaviour. But the relation between mental and behavioural notions is conceptual and hence tighter, and that between mental and neurophysiological notions looser than this proposal allows (Glock, 2009a). Finally, even if the labels and distinctions of natural science are capable of 'carving nature at its joints', in Plato's striking phrase (1997; Phaedrus, 265d–6a), it is a moot question whether our mental labels and distinctions serve such a purpose. And in order to decide that issue, there is no way around establishing what mental terms mean.

Even if one settles for nominal definitions, methodological choices remain. Traditionally, nominal definitions are divided into stipulative definitions on the one hand, and reportive or lexical ones on the other. Stipulative definitions simply lay down *ab novo* what an expression is to mean in a particular context, in complete disregard of any established use it may have. Such definitions cannot be correct or incorrect. But they can be more or less fruitful, in that it may be more or less helpful to single out a particular phenomenon through a separate label. Lexical definitions, at a first approximation, are supposed to capture what an expression does mean in its ordinary use. Note, however, that the term 'ordinary use' is ambiguous. It may refer either to the everyday use of an expression as opposed to its specialist or technical employment, or to the standard use of an expression as opposed to its irregular use,

in whatever area it is employed, specialist employments like those of science included (Ryle, 1971, pp. 301–4). Finally, one should recognize a halfway house between the extremes of unfettered stipulation and faithful articulation of established use: revisionary definitions regiment or modify the extant use of a term, yet without diverging from it completely.

What sort of definition is most appropriate for mental terms, in particular with regard to the question of animal minds? To answer that query, we need to consider, in turn, the use of mental expressions in everyday and specialized contexts.

6.3 Mental idiom in everyday parlance

As regards mental notions, everyday use is both primary and paramount. Making sense of one another in terms of our sensations, feelings, moods, beliefs or desires is not just a cornerstone of our social life but deeply ingrained in the warp and weft of each individual's daily existence. With respect to an idiom that is not just entrenched but pervades our whole lives, unfettered stipulation is rarely advisable. For one thing, it invites confusion for no apparent gain. For another, existing terms, as actually employed, stand in relation to other terms that would have to be redefined as well. This holds not just for the conceptual relations between different mental concepts but also for the connection between the latter on the one hand, concepts from other domains such as moral and legal discourse on the other. Note the contrast with the explanations of mechanical terms in physics. These explanations diverge, often radically, from the everyday understanding. Yet, they do so across the board, clearly and in a precise manner, and one which is universally accepted and patently fruitful. For reasons of principle, there is little prospect of anything analogous with respect to our mental expressions.

Many contemporaries refer to our everyday employment of mental concepts as 'folk psychology'. For the most part, it is further assumed that folk psychology constitutes a scientific or proto-scientific theory of human behaviour. Debates then turn on whether this 'theory':

(1) should eventually be displaced by a more scientific (neurophysiological or computational) theory compatible with physicalism – a more scientific idiom unrelated to our mental notions – even in the case of human beings, as urged by eliminative materialists (e.g. Churchland, 1994).

(2) can be extended to animals, pending the results of empirical disciplines – ethology, neurophysiology, evolutionary biology (e.g. Allen and Bekoff, 1997).

Both debates are predicated on a mistaken presupposition. To be sure, our mental concepts form a systematic network. It does not follow, however, that they constitute a scientific theory, even if the latter includes more than just deductive-nomological theories. Indeed, a conceptual system or vocabulary is categorially distinct from a theory, since it cannot be either true or false (Glock, 2003, p. 256). Of course, statements phrased in terms of our mental vocabulary are truth-apt. Yet, mental notions are not just employed in the third-person, but also in the first-person. And as Wittgenstein emphasized, for the most part, we use psychological sentences in the first-person not to describe, explain or predict anything, but to express or manifest a mental phenomenon (Glock, 1996, pp. 50–4). Third-person uses of mental terms are descriptive, explanatory and predictive rather than expressive. It is arguable, however, that the explanations provided are of a different, non-causal kind (e.g. Alvarez, 2010). Finally, employing mental notions in both the expressive first-personal and the non-causal third-personal way is partly constitutive of being human. If we were to start talking about each other exclusively as bodies steered by neural firings or information processing systems, we would have mutated into a different kind of animal (Bennett and Hacker, 2003, pp. 372–7).

6.4 Mental idiom in philosophy and science

The moral so far: There is no case for abandoning our everyday mental notions or stipulatively redefining them in a more scientific manner. Of course, mental terms are employed not just in everyday parlance but also in philosophy, empirical sciences of the mind and numerous other academic subjects, notably the humanities. These disciplines can and must develop their own terminology and conceptual apparatus. While there is no case for sheer stipulation, there may, nonetheless, be reasons for modifying generally accepted explanations. One might feel, therefore, that for philosophical and scientific purposes we need to graduate from quotidian use towards a more specialized one based on more exacting scrutiny of the phenomena. Thus, one might envisage a 'logical explication' à la Carnap (1956, pp. 7–9). The aim of such an analysis is not to provide a synonym for the analysandum but to replace it by an alternative expression or construction, one which serves the cognitive

purposes of the original equally well while avoiding drawbacks such as obscurity, paradox or excessive ontological commitments.

Now, it may well be both desirable and feasible to modify existing concepts in order to avoid paradoxes or conceptual traps. Yet, for several reasons this does not remove the need for at least starting out with established use.

First, unless the relation between the novel and the established ways of using the pertinent expressions (between the new and the old concepts) is properly understood, the philosophical problems associated with these expressions will merely be swept under the carpet (Strawson, 1963). Secondly, all neologisms and conceptual modifications, those of science included, need to be explained. By pain of regress, this can ultimately be done only in terms of everyday expressions which are already understood – the expressions of a mother tongue. The expressions of our first natural language we acquire not through explanation in terms of another language, but through training in basic linguistic skills. With respect to many specialized purposes ordinary – in the sense of everyday – language is inferior to technical idioms. But it is semantic bedrock. It is only by acquiring ordinary language that we acquire the ability to learn and explain new and technical terms. While 'ordinary language is not the last word ... it is the first word' (Austin, 1970, pp. 103, 185).

Like many philosophical questions, those concerning the mind in general, and animal minds in particular, are phrased in terms of our extant, non-modified vocabulary; indeed, in this case, the idiom is, first and foremost, part of everyday discourse. We would like to know, for example, whether animals can think or are conscious in our sense of these terms, not in a sense introduced by new-fangled philosophical or scientific theories. Answers that employ modified – let alone entirely novel – concepts will simply pass these questions by. They will change the topic or miss it entirely. In fact, unless these modified or novel concepts can be explained coherently through extant concepts, such answers will remain vacuous or obscure.

Our established concepts determine the subject area of most philosophical problems and even of many scientific ones. They are presupposed explicitly or implicitly not just in philosophical theories and arguments, but also in research projects, methods and conclusions from the special sciences. To take an example, the explanation of perception cannot be couched exclusively in everyday concepts but must employ technical concepts from a variety of areas, ranging from psychology to biochemistry. Yet, everyday statements like 'Maria saw that Frank

had put on weight', 'Sarah listens to the Eroica', 'One can smell the wild strawberries' or 'The sense of taste is not affected by old age' pick out the phenomena that the science of perception seeks to explain. Small wonder, then, that in presenting and interpreting the results of empirical research into perception, philosophers and scientists do not uniformly stick to technical terminology. Instead, they often employ everyday terms like 'representation', 'symbol', 'map', 'image', 'information' or 'language' in ways which either remain unexplained or illicitly combine their ordinary uses with technical ones (Bennett and Hacker, 2003; Glock, 2003a).

Notoriously, mental notions give rise to a whole raft of puzzles and perplexities. It is therefore a precondition of any sober approach even to scientific problems involving them that it should pay attention to the established use of the relevant expressions within their normal surroundings. Without the propaedeutic of conceptual clarification, we shall be 'incapable of discussing the matter in any useful way because we have no stable handle on our subject matter' (Joyce, 2006, p. 52).

6.5 Connective analysis

In pursuing any question of the form 'What is X?' we shall inevitably rely on a preliminary notion of X, an idea of what constitutes the topic of our investigation. In our case, we presuppose a preliminary understanding of mental vocabulary. This is not a fully-articulated conception, which would have to emerge from subsequent debates, but an initial idea of what those debates are about. Such a pre-theoretical understanding is embodied in the established uses of the relevant mental terms. In tackling the animal mind and distribution questions, we therefore need to pay heed to our extant mental concepts, as manifested in the standard use of mental terms. Both the explanations of mental concepts and claims about their applicability to animals should be measured, in the first instance, against the established uses of the relevant terms in successful and reasonably controlled forms of discourse. In our case, the latter will include everyday parlance; yet, they will also include specialized disciplines from the behavioural and life sciences, the social sciences and jurisprudence.

Whether such investigations will yield analytic definitions, that is definitions in terms of necessary and sufficient conditions, is another matter. Certain contemporary opponents of conceptual analysis have taken delight in pointing out that ever since Plato, philosophers have failed spectacularly to come up with convincing definitions of any but

the most trivial concepts. Thus, Fodor opines hyperbolically, though not without some license, that 'the number of concepts whose analyses have thus far been determined continues to hover stubbornly around none' (Fodor, 2003, p. 6).

Fortunately, we need not share such extreme pessimism. For one thing, it appears that some central philosophical concepts allow of analytic definitions, once one bids farewell to unjustified assumptions. In other cases, analytic definitions may be in the offing if one takes note of complex ambiguities. Many cases may, indeed, defy analytic definition entirely. This does not mean, however, that it is either impossible or unnecessary to elucidate them. There are other perfectly respectable ways to explain concepts. Nor need these be confined to contextual definitions like the ones Frege gave for numerals, and Russell for definite descriptions, or to the survey of family resemblances provided by Wittgenstein. In this context, Strawson distinguishes between 'reductive' and 'connective analysis' (1992, ch. 2). The former seeks to break down concepts into ultimate components and to unearth the concealed logical structure of propositions. Yet, the later Wittgenstein and Quine have undermined the ideas of ultimate components and definite logical structure. In consequence, connective analysis abandons the analogy to chemical analysis. It is simply the description of the rule-governed use of expressions, and of their connections with other expressions by way of implication, presupposition, and exclusion. Connective analysis need not result in definitions; it can rest content with elucidating features which are constitutive of the concepts under consideration, and with establishing how they bear on philosophical problems and arguments.

Even connective analysis separates conceptual from factual issues, and the elucidation of expressions from the investigation of reality. It is also wedded to the idea that in clarifying conceptual issues we rely not on empirical data, but explicate our understanding of certain terms, drawing in effect on our linguistic competence. Both ideas have been vigorously challenged by Quinean naturalists, who deny that there is any significant difference between analytic or conceptual statements independent of experience and the synthetic or factual statements based on experience. As a result, they maintain that proper or 'scientific philosophy' does not just emulate the methods of the deductive-nomological sciences; it is itself 'continuous with science', and in fact part of science (Quine, 1951, 1970, p. 2).

As I have argued elsewhere, the Quinean attacks on the analytic/synthetic, *a priori/a posteriori* and necessary/contingent distinctions fail. Accordingly, the naturalistic assimilation of the conceptual issues

of philosophy to the factual issues of science is unwarranted (Glock, 2003, chs 2–3). What is correct is that the borders between the conceptual and the factual can and do shift, along with our ways of thinking and speaking. What is more, such changes of the conceptual framework can themselves be motivated by scientific considerations ranging from new experiences and the availability of novel mathematical apparatuses through simplicity and fruitfulness to sheer beauty. But this does not mean that empirical and conceptual propositions are on a par. For such conceptual changes can, in turn, be distinguished from changes of factual beliefs, notably the falsification of scientific theories by empirical evidence.

With respect to specific scientific experiments or lines of reasoning, it is often possible to decide whether or not particular sentences are used empirically or as a definition. The same goes for specific philosophical problems or arguments. Precisely because at present many of these are entwined with scientific issues, it is imperative to disentangle the factual issues ascertained by empirical science from issues of a different kind. These include not just conceptual issues, but also moral and aesthetic ones. It would be futile to pretend, for instance, that rational debate about animal welfare can proceed without distinguishing between conceptual questions (e.g. What should count as a person?), factual questions (e.g. Are there animals satisfying these criteria?), and moral questions (e.g. Can it be legitimate to kill an innocent person?). Behind the backs of their Quinean super-egos, even conscientious naturalists constantly need to draw distinctions of a kind which their official positions disallow or at least cannot account for.

6.6 Impure analysis

In tackling the distribution question, we must pay heed to the conditions for the applicability of mental terms. At the same time, it is obvious that the question to which creatures these terms actually apply also depends on contingent facts about these creatures to be established empirically. This interaction necessitates some qualifications and modifications of the idea that the philosophy of animal minds is simply the conceptual analysis of mental concepts. Indeed, these caveats apply at the very least to all topics on which science and philosophy converge. These are topics that pose both scientific and philosophical problems, and concerning which science therefore gathers fresh data, develops novel methods and constructs new theories. These qualifications are not necessarily incompatible with the kind of conceptual analysis inspired

by Wittgenstein and Ryle, but they have not been sufficiently recognized in these quarters. Furthermore, they stand in tension with the received image of conceptual analysis as a purely *a priori* exercise unaffected in all respects by scientific findings. Accordingly, they lend succour to a type of conceptual analysis that I call 'impure'.

The first caveat is this: While conceptual and factual problems are distinct in principle even when it comes to topics on which philosophical puzzles and scientific theory formation converge, they cannot be tackled in isolation (similarly Dupré, 2009, pp. 244–9). More generally, philosophers cannot engage in a second-order conceptual reflection on a mode of discourse without at least some acquaintance with its first-order problems, claims and methods. Philosophical problems about X cannot simply be resolved by empirical discoveries and theories about X. Yet, they cannot be resolved without taking note of the latter, since they often determine what the conceptual issues are, and legitimately so.

A second caveat. On the one hand, matters of meaning antecede matters of fact. As emphasized above, it makes sense to investigate a phenomenon X only if it is clear what is to count as X, if only provisionally. On the other hand, we must avoid the Socratic mistake of thinking that one cannot establish empirical facts about X unless one already has an analytic definition of 'X'. In Plato's *Meno* (1997, 80 a–e), Socrates devises the following paradox. It is impossible to inquire into what X is, since one cannot look for or recognize the correct answer, without already knowing it from the start. The underlying argument runs roughly as follows:

P_1 To recognize the correct definition of X we already have to know what X is.
P_2 To know what X is, is to know the definition of X.
C_1 We would already need to know the correct definition of X in order to recognize it.
C_2 The search for a correct definition of X is pointless.

P_1 is mistaken, at least in conjunction with P_2, which identifies knowing what X is with knowing a definition of X. As Kant pointed out, definition marks at best the terminus of philosophical inquiry, not its beginning. And as Wittgenstein pointed out, to look for and recognize the correct explanation of X, all one needs is a pre-theoretical understanding of 'X', something we acquiring mastering the use of terms expressing the concept of X.

Any theory of mental properties in cognitive science presupposes at least a certain preconception of what counts as a mental property. But this does not mean that one needs a cast-iron, precise definition of these properties in advance of empirical theory building, contrary to Socrates. Our concepts are tools which we fashion for our purposes, in science the purpose of describing, explaining and predicting phenomena. In scientific theory building, definitions are to be read from right to left: We introduce labels for newly discovered or postulated phenomena. Scientists do not, by and large, first define a notion and only then consider to what, if anything, it applies to. An example from ethology. Tomasello and Call (1997, ch. 9) differentiate types of 'social learning', such as 'emulative learning' and 'imitation'. These categories were not devised in the armchair and only subsequently applied to observable phenomena; rather, they are based on observing the interactions between apes confronted with novel situations and comparing them with those between children.

Thirdly, our pre-theoretical mental concepts are not just complex but often vague and polysemous. As a result, these concepts may leave greater leeway when it comes to deciding whether or not they are, in principle, applicable to animals of various kinds than was traditionally assumed. To that extent, the impact of conceptual reflection on the problem of animal minds decreases, while that of empirical research increases. Novel scientific findings may sway us to draw the line in a more determinate manner than previously would have been apposite – with the proviso that there is something about our extant concepts that allows us to draw the line in this novel and more precise fashion.

A fourth and connected caveat: The conceptual connection between mental and linguistic capacities is less tight and far-reaching than most conceptual analysts have maintained or assumed. Creatures that resemble us very closely in all respects apart from language – as regards not just intelligent behaviour like tool-manufacture but also facial expressions and bodily demeanour – can, in principle, be credited even with rudimentary reasoning capacities (Glock, 2009a).

Our closest evolutionary ancestors without language probably resembled us in many of these respects. Yet, at present, there are no creatures that are sufficiently close to us in all respects bar language for this consideration to get a grip. Given this contingent fact, the philosophy of animal minds needs to consider the extent to which the connection between mind and language holds, given the actual capacities of extant species. For instance, ascribing beliefs of a certain complexity may presuppose their linguistic manifestability, given the absence of other

behavioural manifestations accessible to us. But it obviously an empirical question what spectrum of non-linguistic manifestations features in extant animal species.

This interaction and mutual dependence of conceptual and factual considerations should neither come as a surprise nor be regarded as a threat to conceptual analysis. Wittgenstein, for one, distinguished not just factual propositions and the conceptual rules that underlie them, he allowed that empirical discoveries could motivate conceptual change. He further pointed out that there is a 'framework' of contingent facts concerning our own nature and the world around us without which a certain conceptual apparatus or an entire conceptual scheme may be pointless or even unfeasible (Glock, 1996, pp. 135–9). It is almost inevitable that certain empirical facts come into play once we consider the question of whether certain abilities – such as the ability to entertain beliefs and desires – presuppose certain other abilities – such as the ability to express such beliefs and desires in sentences. For whether one ability requires another, will commonly depend on what other abilities are being assumed. And what abilities can be assumed with respect to a biological species – whether extant or extinct – is once more an empirical issue.

Impure conceptual analysis contrasts with several widespread trends in the philosophy of animal minds. One is a naïve naturalism that simply accepts at face value the claims made by cognitive scientists without scrutinizing their conceptual and methodological credentials. A second is Quinean naturalism, which purports to provide a sophisticated semantic rationale for such a simple-minded procedure. A third alternative seeks to make the question of animal minds more tractable by modifying extant mental concepts. There is nothing wrong with this in principle. However, such empirically or theoretically inspired concept-formations must be accompanied by a reflection on the – possibly provisional – understanding of the concepts that has informed specific theories, experiments or lines of research so far. Otherwise, the theories may simply miss their purported topic. They may, indeed, solve a tractable problem, yet it might not be the one we were originally pursuing. And although we might abandon the original problem in favour of the new one, we need to be clear about how they are connected.

Mutatis mutandis for eliminativism with respect to animal minds, at least for the purposes of science and philosophy. Why not react to the difficulty of 'grand' questions like whether animals can think or act for reasons by simply dropping them and sticking to scientific observations

about what animals are capable of doing and neurophysiological explanations of these exploits? Some influential philosophers of science notwithstanding, there are, indeed, observations that are not theoretically laden. Nevertheless, any scientific observation will be conceptually laden. Although some statements may be free from theoretical commitments, there is no such thing as a non-conceptual statement. Furthermore, as the failure of behaviourism shows, concepts suitable for describing animal behaviour will either be completely uninformative or mentalistic, at least implicitly. Even austere concepts like stimulus and response are contested. In any event, we would still need to relate the envisaged empirical findings expressed in non-mentalistic vocabulary to our original 'mentalistic' questions. For it is these questions that link up with others on which we do need to take a stance, notably moral questions concerning our treatment of animals (Glock, 2012).

Among some naturalists, there is a tendency to reduce the philosophy of animal minds to the philosophy of cognitive ethology. As we have seen, however, mental notions are rooted in non-scientific discourse. A proper account of them must take note of this core area and its logico-semantic analysis. It cannot make do with methodological and conceptual reflections on research strategies in the cognitive sciences.

Another naturalistic approach may actually welcome the analysis of the established patterns of applying mental terms to animals in everyday life and science. But it goes beyond it, in the name of scientific theory-construction. Clarifying the extent to which mental notions can meaningfully be applied to animals is not sufficient, its proponents insist. It is imperative to provide an explanatory theory of how it is possible for animals to possess the mental capacities that we correctly credit them with. In particular, it is not enough to show that animadversions to the idea of intentional states in animals misconstrue notions like that of belief. One must also explain how intentionality or 'mental representation' is possible, given naturalistic constraints.

Impure conceptual analysis has no qualm with granting that a causal explanation of the genesis of mental capacities must be possible, and that any analysis which rules out the possibility of such an explanation must be deficient. At the same time providing such a genetic theory is the business of empirical disciplines, namely those investigating the ontogenetic and phylogenetic development of mentally gifted animals. Consolation from philosophy, by contrast, is both required and in store when the 'How is it possible?' question is rooted in a conceptual problem. Thus, the idea that there is anything mysterious about the very possibility of consciousness or thought in natural creatures – human or

non-human – is based on a confused conception of these phenomena and of what may count as natural.

Advocates of an explanatory theory typically pursue a bottom-up strategy. They note or envisage organic and inorganic systems of increasing complexity. These are then compared and contrasted with both animals and humans, mostly in order to show that the differences among all three cases are less significant that commonly supposed (Millikan, 1984; Dretske, 2000, ch. 12). Now, finding and inventing objects of comparison is a valuable tool of philosophical clarification. Yet, such comparisons yield insights only if the terms they employ are used in cognisance of their established meaning in application to the respective comparata.

Finally a challenge from the opposite direction, that of pure conceptual analysis. Should we not simply avoid the empirical issues by extracting purely conceptual questions about animal minds? For instance: Does it make linguistic sense to apply mental expressions to non-linguistic creatures? Would anything count as non-linguistic creatures thinking that something is the case? However, even if such an approach can overcome the hurdles mentioned in my caveats, it is difficult to see how one might make progress with these questions without at least considering real animal behaviour as a heuristic device. Furthermore, such a pure approach is barren in at least one respect. Even philosophers have been interested in animal minds in the context of contemplating what mental capacities they actually have. Unsurprisingly, since it is that issue which has numerous important implications inside and outside of philosophy. Even when animals are considered solely for the purposes of establishing the contours of our mental notions, they can only serve that role if one takes note of their actions and capacities. Empirical findings without conceptual reflections are blind; conceptual reflections without empirical findings are empty – they leave unresolved questions that motivate even most philosophical inquiries into animal minds.

6.7 Methodological principles between philosophy and science

There is a further respect in which the purity of conceptual analysis must be sacrificed, especially concerning animal minds but also more widely in the philosophy of psychology. While the distinction between conceptual (*a priori*) and factual (empirical) questions and statements is legitimate and important, it is not exhaustive, even leaving aside

normative moral principles. There is a sphere of methodological considerations that straddles or sits uneasily between the two.

Contrary to verificationism, the idea of a 'method of verification' does not afford a general direct connection between meaning (concepts) and methodology. Under what conditions a term is applicable to something is part of its meaning. But how the (non-) application of a term is to be verified or falsified is not necessarily part of its meaning. For it may depend on factual considerations of either a specific or theoretical kind. Even if there is a link between meaning and verification, not all aspects of the method for establishing the truth-value of a proposition are relevant to its meaning, but only those which must be known to qualify as a competent speaker. Thus, it is wrong to suggest that the fact that we can learn about who won the boat race by reading a newspaper goes some way to explaining the meaning of 'boat race'. Similarly, that the length of playing fields is measured through the use of tripods is a matter of physics, while to say that measuring involves the possibility of comparing the lengths of different objects is partly constitutive of the meaning of 'length' (Glock, 1996, pp. 382–5).

Turning to the investigation of animals, one methodological issue concerns the respective merits of experiment and observation. Should we set more store in field observations or controlled experiments? The latter allow of more reliable corroboration and of systematically alternating the parameters of the situation. The former are more significant for biological purposes, notably those of evolutionary theory and ecology. They possess greater 'ecological validity', to use a term from research design.

These are not straightforwardly empirical matters, since they concern what kind of empirical evidence should carry what kind of weight. Nor are they straightforwardly issues of a conceptual kind. It is not part of the meaning of 'mind' or 'behaviour', for instance, that behaviour observed under natural conditions should reveal more about a subject's mental capacities than behaviour elicited as part of an experiment. Furthermore, within practical constraints ethologists can aspire to the best of both worlds by employing modern technology in order to control for the relevant parameters even in the wild (e.g. Cheney and Seyfarth, 2007).

Nonetheless, the contrast carries a potential for puzzles and quandaries with a philosophical dimension. For one thing, while atypical behaviour by a specimen – for example symbol use by enculturated bonobos under experimental conditions – clearly evinces mental capacities, it is far from clear what the presence of these capacities under those conditions

shows about the nature of the species, or about the proximity between bonobos and us. For another, a basic methodological dilemma looms.

On the one hand, the more unrestricted and spontaneous animal behaviour, the less rigorous the procedure and the more it relies on 'mere anecdotes'. There is also the danger of the notorious 'Clever Hans effect'. We need to exercise caution in interpreting experiments in which animals interact with humans, since the latter may unwittingly aid the animals in performing a desired task, by conditioning them to respond to unconscious human signals.

On the other hand, the more controlled and predictable animal behaviour, the more artificial, and hence less ecologically sound, the findings. Thus, the symbolic systems acquired by enculturated apes are remote from their systems of communication in the wild. Furthermore, rigorous procedures such as duplication or 'double-blind strategies' to protect against the Clever Hans effect may simply undermine the subject's willingness to cooperate (Dupré, 2002, ch. 11). Yet, the less restricted and spontaneous animal behaviour, the less rigorous is the procedure, and the more it relies on 'anecdotal evidence'.

6.8 Morgan's canon

Another hot potato is 'Morgan's canon': 'in no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale' (Morgan, 1894, p. 53).

Although it is a well-known methodological principle of comparative psychology, it is far from obvious what Morgan's canon actually entreats us to do, and on what grounds. It is often regarded as an instance of a general principle of parsimony, namely 'Occam's razor'. And that principle is held to hold for all academic subjects. Alas, it is unclear what kind of economy is at issue. It is even less clear how that parsimony relates to other desiderata of theories, such as explanatory power, simplicity, conservatism, modesty, precision, facility of computation and avoidance of perplexities. As a result, Morgan's canon lends itself to competing interpretations and to exploitation by different, sometimes diametrically opposed positions.

In some versions, Morgan's canon is a principle less of economy than of conservatism. Morgan himself veered towards this kind of gloss in the second edition of his book. There he added 'that the canon by no means excludes the interpretation of a particular activity in terms of the higher processes if we already have independent evidence of the occurrence of

these higher processes in the animal under observation' (1894, p. 59). The point here is not that we should make do with as little as possible; it is rather that we should add as little as possible over and above the phenomena already invoked.

Next, Morgan's canon is not simply a demand for simplicity. It expresses a preference concerning kinds of parameters rather than numbers of parameters. Many ethological debates pit 'boosters', who explain the often astonishing accomplishments of animals by crediting them with a wide variety of mental capacities, against 'scoffers', who try to make do with as few as possible. While the latter are prone to invoke Morgan's canon against the appeal to higher mental phenomena, they are downright profligate when it comes to the number of less advanced mechanisms they detect, whether it be purely physiological mechanisms to explain, for example, discrimination or feats of associative learning to explain, for example, transitive inference and 'mind-reading'. Such an insistence on psychological simplicity amounts to the idea that the complete denial of animal mentality constitutes a kind of null-hypothesis, to be abandoned only if the data compel it.

On the other hand, one can turn the table on the scoffers by appeal to a different kind of parsimony, one connected to the desideratum of explanatory power. Here, the virtue upheld is one of psychological unity. Everything else being equal, we should seek a unified approach to animal and human behaviour, one that subsumes as many behavioural phenomena as possible under a single explanation. A more specific version of this line of reasoning invokes evolutionary parsimony. If closely-related species like humans and great apes behave in similar ways, it is simpler, from an evolutionary point of view, to posit homologous mental mechanisms, rather than to assume that analogous mechanisms evolved independently (Sober, 2009, pp. 250–6).

In both versions, Morgan's canon turns into a weapon in the arsenal of boosters. And both are intimately connected to an argument from analogy: The simplest explanation is the one which explains similar behavioural outcomes as effects of similar causes, whether they be proximate mental causes or ultimate evolutionary ones. Arguments from analogy are themselves tricky and controversial. However, one need not accept booster arguments fuelled by simplicity and analogy considerations to feel that Morgan's canon is heavy artillery indeed. I propose to replace it by something more modest. Call it Glock's canon if you please, even though it is in fact more like a handgun. We should only attribute higher mental capacities to a creature if this is the best explanation of its behavioural capacities. Like Morgan's canon, this modified principle

relies on a (gradual) classification of mental capacities into higher and lower, which requires explanation and defence. Both can be supplied, I hope, by combining philosophical demarcations among perception, understanding and reason with congenial distinctions in contemporary cognitive ethology.

Weakening Morgan's principle in the way here proposed would put paid to a widespread (mal)practice among sceptics about animal minds. It is common to account for the cognitive achievements of apes and cetaceans by invoking more or less far-fetched feats of associative learning, simply in order to scotch the prima facie compelling suggestion that these animals engage in genuine planning or reasoning (e.g. Povinelli and Vonk, 2006; cf. Tomasello and Call, 2006). Indeed, some sceptics seem willing to postulate mechanisms for which there is no evidence whatever, and which are perhaps not even coherently describable, for the sole purpose of avoiding the attribution of higher mental faculties.

In any event, the controversies surrounding Morgan's canon once more defy a neat classification into the conceptual and the factual, or into the philosophical and scientific. Ontological principles of parsimony indisputably raise philosophical problems. Yet, not all of these are conceptual in a straightforward sense, as debates in the philosophy of science amply demonstrate. And as regards animal minds, at least, both the conceptual and methodological issues discussed by philosophers are intricately intertwined with scientific issues of both a methodological and a theoretical kind.

6.9 Animals as touchstones

Let me end with one other area that is central to the question of how the philosophy of animal minds ought to be conducted. It is the interaction between the philosophical elucidation of mental notions and the fact that we regularly apply such notions to animals in both everyday life and science. On the one hand, the applicability of mental notions to animals depends on how they are to be construed. On the other hand, animals can function as a test-case for the analysis of a significant number of mental concepts. This idea was already mooted by Hume, who maintained that animals are 'a kind of touchstone, by which we may try every system in this species of philosophy' (1978, 1.3.16). Hume employed this 'animal test' both positively and negatively (Wild, 2006, pp. 270–2). He maintained that his empiricist explanations of human mental life by reference to a few simple mechanisms enjoy a crucial

advantage: they are applicable to animals as well, thus providing a uniform account. Conversely, theories which explain our mental life by reference to rational capacities unattainable for beasts have a smaller scope. Furthermore, they founder because they fail to explain the same mental operations through the same mechanisms. Indeed, Hume boasted that the animal test furnishes an 'invincible proof' of his system and a 'clear proof of the falsehood' of rationalist alternatives (1978, 1.34.16).

'Is not this fine reasoning?' One is inclined to echo Hume's sardonic comment about moral rationalism (1975, Apd. I). Except that the circularity evidenced in his passage is more blatant than the one he diagnosed in moral rationalism. Wielding the animal test as a weapon against rationalist accounts of human mentality presupposes what it seeks to establish, namely that even an apparently sophisticated 'mental operation' can be 'common to men and beasts'.

On a more positive note, Hume's warning against overblown pictures of human mentality is well taken. Many denials of mental phenomena in animals are based on an overly intellectualist picture of what these phenomena amount to in humans. Furthermore, there may yet be a version of the animal test that is acceptable.

A refusal to apply even simple mental notions to animals may necessitate a novel construal of mental notions. Such revisionism faces a serious challenge. If mental concepts are applied to animals in a coherent and well-controlled manner in established forms of discourse (whether everyday or scientific), then any construal which rules out such application will fail to capture these notions, at least as employed in these areas.

At the same time, those who deny that there are significant differences between us and animals often ignore that the paradigmatic instances of most mental properties are adult, linguistic humans. To disregard core cases in favour of marginal, exotic, limiting or contested cases is, alas, a long-standing tendency within philosophy at large. Nevertheless, it is a case of the tail wagging the dog. Our mental concepts should not be construed exclusively, or even primarily, by reference to their application to animals.

These discussions of the impurity of conceptual analysis reinforce, rather than diminish, the need for conceptual and methodological elucidation. While philosophy is neither part of, nor continuous with, science, it has a contribution to make. Logic and mathematics are neither part of, nor simply continuous with, empirical science. Nonetheless they contribute to the latter by providing the formal methods of proof and calculation that are essential to empirical science. Philosophy can

aid science in a different manner, namely by obviating confusions that lie in its path. With respect to the sciences, therefore, philosophy is no longer the queen (as in Aristotelianism). It is not even primarily the judge who holds the sciences accountable to standards of knowledge (as in Kantianism) or linguistic sense (as in Wittgensteinianism). Its role is more akin to that of the Lockean underlabourer. Philosophical reflection on topics successfully investigated by empirical disciplines should not just be conceptually enlightening and methodologically scrupulous, but also beneficial to empirical research. At the same time, conceptual underlabour is both indispensible and tricky. Philosophers ought not to shirk it by exclusively dabbling in science (let alone pseudo-science) instead, and scientists should be duly grateful.

References

C. Allen (2010) 'Animal Consciousness', *The Stanford Encyclopaedia of Philosophy*, http://www.science.uva.nl/~seop/entries/consciousness-animal/.

C. Allen and M. Bekoff (1994) 'Intentionality, Social Play and Definition', *Biology and Philosophy*, 9(1), 63–74.

C. Allen and M. Bekoff (1997) *Species of Mind* (Cambridge/Mass.: MIT Press).

M. Alvarez (2010) *Kinds of Reasons: An Essay in the Philosophy of Action* (Oxford: Oxford University Press).

J. L. Austin (1970) *Philosophical Papers* (Oxford: Oxford University Press).

M. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).

R. Carnap (1956) *Meaning and Necessity* (Chicago: University of Chicago Press).

D. L. Cheney and R. M. Seyfarth (2007) *Baboon Metaphysics* (Chicago: University of Chicago Press).

P. M. Churchland (1994) 'Folk Psychology (2)' in S. Guttenplan (ed.) *A Companion to the Philosophy of Mind* (Oxford: Blackwell), 308–16.

F. Dretske (2000) *Perception, Knowledge and Belief* (Cambridge: Cambridge University Press).

J. Dupré (2002) *Humans and Other Animals* (Oxford: Oxford University Press).

—— (2009) 'Hard and Easy Questions about Consciousness' in H. J. Glock and J. Hyman (eds) *Wittgenstein and Analytic Philosophy* (Oxford: Oxford University Press), 244–49.

J. Fodor (2003) *Hume Variations* (Oxford: Clarendon).

H. J. Glock (1996) *A Wittgenstein Dictionary* (Oxford: Blackwell).

—— (2003) *Quine and Davidson on Language, Thought and Reality* (Cambridge: Cambridge University Press).

—— (2003a) 'Neural Representationalism', *Facta Philosophica*, 5(1), 147–71.

—— (2008) *What Is Analytic Philosophy?* (Cambridge: Cambridge University Press).

—— (2009a) 'Can Animals Act for Reasons?' *Inquiry*, 52(3), 232–55.

—— (2009b) 'From Armchair to Reality? (Timothy Williamson's Philosophy of Philosophy)', *Ratio (2010)*, XXIII(3), 339–48.

—— (2012) 'Mental Capacities and Animal Ethics' in K. Petrus and M. Wild (eds) *Animal Minds and Animal Ethics* (New York: Springer).
P. M. S. Hacker (1996) *Wittgenstein's Place in Twentieth-Century Analytic Philosophy* (Oxford: Blackwell).
O. Hanfling (2000) *Philosophy and Ordinary Language* (New York: Routledge).
D. Hume (1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (Oxford: Oxford University Press).
—— (1978) *A Treatise of Human Nature* (Oxford: Oxford University Press).
R. Joyce (2006) *The Evolution of Morality* (Cambridge, MA: MIT Press).
S. Kripke (1980) *Naming and Necessity* (Cambridge, MA: Harvard University Press).
J. Locke (1975) *An Essay Concerning Human Understanding* (Oxford: Oxford University Press).
R. G. Millikan (1984) *Language, Thought and other Biological Categories* (Cambridge, MA: MIT Press).
C. L. Morgan (1894) *An Introduction to Comparative Psychology* (London: Walter Scott).
Plato (1997) *Plato: Complete Works* (Indianapolis: Hackett).
D. Povinelli and J. Vonk (2006) 'We Don't Need a Microscope to Explore the Chimpanzee's Mind' in S. Hurley and M. Nudds (eds) *Rational Animals?* (Oxford: Oxford University Press), 385–412.
H. Putnam (1975) *Mind, Language and Reality* (Cambridge: Cambridge University Press).
W. V. Quine (1951) 'Two Dogmas of Empiricism', *From a Logical Point of View* (Cambridge, MA.: Harvard University Press), 20–46.
—— (1970) 'Philosophical Progress in Language Theory', *Metaphilosophy*, 1(1).
G. Ryle (1971) *Collected Essays*, Vol. II (London: Hutchinson).
E. Sober (2009) 'Parsimony and Models of Animal Mind' in Robert W. Lurz (ed.) *The Philosophy of Animal Minds* (Cambridge: Cambridge University Press) 250–6.
P. F. Strawson (1963) 'Carnap's Views on Constructed Systems vs. Natural Languages in Analytic Philosophy' in P. A. Schilpp (ed.) *The Philosophy of Rudolf Carnap, Library of Living Philosophers* (La Salle: Open Court), 503–18.
—— (1992) *Analysis and Metaphysics* (Oxford: Oxford University Press).
M. Tomasello and J. Call (1997) *Primate Cognition* (Oxford: University Press).
—— (2006) 'Do Chimpanzees Know What Others See – or Only What they are Looking At?' In S. Hurley and M. Nudds (eds) *Rational Animals?* (Oxford: Oxford University Press), 371–84.
A. White (1987) *Methods of Metaphysics* (London: RKP).
M. Wild (2006) *Die Anthropologische Differenz* (Berlin: de Gruyter).

7

Realism, but Not Empiricism: Wittgenstein versus Searle

Danièle Moyal-Sharrock

7.1 Wittgenstein: realism without empiricism[1]

In a note on page 56 of the *Philosophical Investigations*, Wittgenstein writes:

> What we have to mention in order to explain the significance, I mean the importance, of a concept, are often *extremely* general facts of nature: such facts as are hardly ever mentioned because of their great generality. (PI, p. 56, bottom note, my emphasis)

And in Part II of the *Investigations* – or what, according to Peter Hacker and Joachim Schulte's new translation, is not really part of the *Investigations* at all, but a 'fragment' of his philosophy of psychology (and I concur) – Wittgenstein writes:

> If the formation of concepts can be explained by facts of nature, should we not be interested, not in grammar, but rather in **that in nature which is the basis of grammar**? – Our interest certainly *includes* the correspondence between concepts and very general facts of nature. (Such facts as mostly do not strike us because of their generality.) But our interest *does not fall back upon* these possible causes of the formation of concepts; we are not doing natural science; nor yet natural history – since we can also invent fictitious natural history for our purposes.[2] (PI II, p. 230, xii; bolded emphasis mine, italicized original)

The variant of this passage found in the *Remarks on the Philosophy of Psychology*, Part I is, in some sense, clearer; here it is:

> If we can find a ground for the structures of concepts among the facts of nature (psychological and physical), then isn't the description of the structure of our concepts really disguised natural science; ought we not in that case to concern ourselves not with grammar, but with what lies at the bottom of grammar in nature?
>
> Indeed **the correspondence between our grammar and general (seldom mentioned) facts of nature** does concern us. But our interest does not fall back on these *possible* causes. We are not pursuing a natural science; our aim is not to predict anything. Nor natural history either, for we invent facts of natural history for our own purposes. (cf. PI, p. 230a; RPP I, §46, bolded emphasis mine, italicized original)

This[3] makes it clear that when Wittgenstein speaks of the correspondence between concepts and nature, he is talking about the correspondence between the *structures* of concepts – that is, our grammatical rules – our grammar – and facts of nature. Take the concept of pain, some of the 'structures' of that concept can be expressed in such grammatical rules as: 'Human beings are normally susceptible to pain,' 'Tables and chairs don't feel pain,' and 'There is psychological as well as physical pain'. In these passages, then, Wittgenstein is saying that, of course, we are interested in the correspondence between our concepts (or our grammar, as the latter version has it) and very general facts of nature – 'Indeed – he writes in *On Certainty* – does it not seem obvious that the *possibility* of a language-game is conditioned by certain facts?' (OC, §617, my emphasis; see also RFM, §§80, 116). Yet, it is not as natural scientists or natural historians that we are interested in this correspondence, but as philosophers; that is, we are interested in the logical, not in any empirical or otherwise explanatory nature[4] of this correspondence: 'our interest does not fall back on these *possible* ['possible' italicized in the original] causes.' And so, these very general, seldom mentioned, facts of nature will be *included* in any perspicuous presentation of concept-formation, but not as objects of empirical hypotheses; rather, as 'given', as 'what has to be accepted' (PI II, p. 226). For philosophy does not take its impetus or material from empiricism or empirical enquiry, but nor, however, is it detached from the world: indeed, Wittgenstein urges us to not think but to look (PI, §66). And, by this, he means that the *object* of philosophical enquiry will always be human practices, human life; and its *method* of enquiry, a non-scientific way of observing human life and making perspicuous presentations about it. Granted, philosophy is concerned with conceptual elucidation, but our concepts are inextricably related to our life.

And this is why, to the question 'Could a legislator abolish the concept of pain?' Wittgenstein's reply is negative: 'The basic concepts are interwoven so closely with what is most fundamental in our way of living that they are therefore unassailable' (LW II, p. 43–4).

What does it mean for a concept to be unassailable? It means for it to be nonhypothetical, unfalsifiable, indubitable – but not in the sense of having been proved beyond any doubt to be a 'fitting' description of reality, rather, in the sense of not being susceptible of doubt at all. 'Unassailable' means 'grammatical' or 'logical'. The correspondence between our basic concepts and 'very general facts of nature', or 'what is most fundamental in our way of living', is a logical, not an empirical or a rational correspondence. Let me give this a little more backing with the help of *On Certainty*.

Here, Wittgenstein is concerned with our basic certainties, which include certainties such as 'Human beings need oxygen, nourishment, sleep; they can feel pain, use language, think etc.' – and so, certainties that *correspond to* those very general facts of nature he was talking about. But how *do* they correspond? He is clear that we do not come to our basic certainties from reasoning: 'I cannot say that I have *good grounds* for the opinion that cats do not grow on trees or that I had a father and a mother' (OC, §282, my emphasis). We did not get to the certainty: 'Humans think' from having observed humans, chairs and tables and concluded from these observations that human beings can, while chairs and tables cannot, think. While some of our basic certainties may have their root in experience, they are not *inferred* from experience, or from the success or reliability of some experiences. Of course, it *is* the case that natural phenomena are reliable – that we can 'count on' human beings needing oxygen and mountains not sprouting up in a day – yet, this certainty is not *derived from, justified by*, or *grounded in* any natural or empirical regularity:

> No, experience is not the *ground*[5] for our game of judging. Nor is its outstanding success. (OC, §131, my emphasis)

And here it is worthwhile noting (though we will come back to it) that their not being the product of ratiocination is one of the features of basic certainties that prompts Wittgenstein to realize that, although they often have the 'form of' empirical propositions, our basic certainties are *not* empirical propositions but rules of grammar (OC, §401).

Now, what Wittgenstein also realizes is that although our basic certainties/rules of grammar[6] are not *rationally* grounded in recurrent

experience and success, they can be *caused* by recurrent experience and success:

> This game proves its worth. That may be the *cause* of its being played, but it is not the *ground*. (OC, §474, my emphasis; cf. also OC, §§130, 429)

The only correspondence between reality and grammar is *causal*, but causal in the sense of conditioned (as opposed to reasoned):

> Certain events would put me into a position in which I could not go on with the old language-game any further. In which I was torn away from the *sureness* of the game.
>
> Indeed, doesn't it seem obvious that the possibility of a language-game is conditioned by certain facts? (OC, §617)

Though not *justified* by facts, our grammar is *conditioned* by facts; by the world we live in – indeed by regularities in the world. Let me try to explain.

A fact may be at the origin of a grammatical rule, but it will have been transformed into a rule as a result of conditioning, not reasoning. Repeated exposure will have, as it were, hammered the fact into the foundations of our thought:

> We say we know that water boils and does not freeze under such-and-such circumstances. Is it conceivable that we are wrong? Wouldn't a mistake topple all judgment with it? More: what could stand if that were to fall? Might someone discover something that made us say 'It was a mistake'?
>
> Whatever may happen in the future, however water may behave in the future, – we *know* that up to now it has behaved *thus* in innumerable instances.
>
> This fact is fused [*eingegossen*] into the foundations of our language-game. (OC, §558, emphasis original)

Some facts have been fused into the bedrock, have become part of our conceptual scaffolding. Wittgenstein's image of *a fact* being *fused into* (or infused, or cast in, or poured into: *eingegossen*) our foundations is deliberate and crucial. It reminds us that many conceptual necessities are related to facts (or *a posteriori* discoveries), but that these facts have become part of our foundational or grammatical bedrock *through a non-epistemic process* – like repeated exposure or training (although our initial

awareness of them might have been epistemic or empirical). Essentially, in the last sentence of the above passage, Wittgenstein is saying, before Kripke, that some of our conceptual necessities have their origin in *a posteriori* discoveries. As Rom Harré and Edward Madden explain:

> It is contingent that any man is a father, but conceptually necessary that being a father he has (or has had) a child. But that conceptual necessity is a reflection of the natural necessity of the father's role in the reproductive process, a role not known to some Aboriginal tribes even in historical times... The conceptual necessity has come into being in response to an *a posteriori* discovery of the natural necessity of the father's role.... But so deeply has this conceptual necessity become embedded in the language, we forget that it has its source in an *a posteriori* discovery. (Harré and Madden, 1975, p. 48)

A rule of grammar is not *answerable* to reality, but is assumed in all our language-games about reality; in all we can say or ask about reality. It is part of the unquestioned framework that allows us to form our questions and hypotheses about reality: 'it is anchored in all my questions and answers, so anchored that I cannot touch it' (OC, §103). This makes our expectation – if one can call it that – that 'Humans can experience pain' or 'Babies do not kill' as unreasoned and automatic an expectation as that the sum of '2+2' is '4' (cf. OC, §§448, 653). We mistakenly think that we come to the grammatical rule (or what most philosophers call 'the basic belief'): 'Babies do not kill' in the same way we come to a conclusion from reasoning. This confusion is due to our always assuming that some reasoning, inference, rationalization, justification had taken place where there had, in fact, been none. Grammar is not the result of observation or of thought; it constitutes the 'scaffolding' of thought[7] (OC, §211).

When Wittgenstein writes that 'The common behaviour of mankind is the *system of reference* by means of which we interpret an unknown language' (PI, §206, my emphasis), he means that it is this shared human behaviour (what he elsewhere also refers to as our 'steady ways of living, regular ways of acting', PO, §397) – such as that human beings experience pain and joy and sadness; that they need to eat, sleep, breathe air; that they have the visual apparatus they do – that constitutes the bedrock from which *any human being* can begin to understand a foreign language. These 'very general facts of nature' – the common behaviour of mankind, as well, of course, as those very general or basic facts of our natural world that do not only concern human beings – such as that

apples fall from trees, and cows do not grow on them; that mountains do not sprout up in a day and the world has existed for a very long time – make up a universal 'system of reference' from which all human languages get their grammar.

What Wittgenstein is here implicitly suggesting is that there is a universal grammar at work here, not in the Chomskyan sense – not a set of linguistic principles embedded in the brain, but a bedrock of grammatical rules that act as bounds of sense for any normal human being. Grammatical rules such as 'Human beings can tell stories'; 'If you cut someone's head, the person will be dead' – these, and not any linguistic principles – make up the bedrock or universal grammar from which any human being can start to interpret a foreign language. Those facts, Wittgenstein would say, *belong to grammar*, for as he writes: 'What belongs to grammar are all the conditions (the method) ... necessary for the understanding (of the sense)' (PG, p. 88). And he also writes:

> I will count any *fact* whose obtaining is a presupposition of a proposition's making sense as belonging *to language*. (PR, §45, my emphasis)
>
> What looks as if it *had to exist*, is part of the language. It is a paradigm in our language-game; something with which comparison is made. And this ... is ... an observation concerning our language-game – our method of representation. (PI, §50, my emphasis)

So that sentences that express such 'very general facts of nature' are not descriptions, but a 'framework for description' (RFM, §356). And to utter the sentence: 'Human babies can look after themselves' is not to give a wrong description of the facts or to make a wrong empirical statement, but to utter nonsense – to transgress the bounds of sense. That babies cannot look after themselves is part of the universal grammar of any human language: it is one of those grammatical rules or certainties that underlie 'all questions and all thinking'[8] (OC, §415).

In being concerned with *conceptual* elucidation, then, philosophy is inextricably concerned with our life, though not *empirically* concerned. Empiricism is, for philosophy, off-limits. It is not to be ignored, but nor is it to be used. Wittgenstein's realism is not a form of empiricism – and here I am using the term 'realism' as it was used before its association with materialism, as having to do with life, reality. In the *Remarks on the Foundations of Mathematics*, Wittgenstein writes: 'Not empiricism and yet realism in philosophy, that is the hardest thing' (RFM, §23). Philosophy marks the limits of the empirical. And this, in turn, is what marks the *autonomy* of philosophy, as, relatedly, the autonomy

of *grammar*. And, as we have just seen, grammar's being autonomous (PG, §63) does not mean that it has no link with reality, but that it is not '*answerable* to any reality' or '*accountable to* any reality' (BT, §184; PG, §184). We saw that what Wittgenstein means by this is that grammatical rules are not *rationally justified* by reference to anything empirical.[9] The relationship between grammatical rules and reality is not a rational one; we can neither *justify* nor *invalidate* a grammatical rule empirically.

To say that grammar is not *justifiable* by empirical facts or established by human decision is also to recognize its objectivity. If grammatical rules are the product of agreement (RFM, §353; Z, §§428–30), it is not a concerted or deliberate agreement, but what Wittgenstein calls a 'peaceful agreement' (RFM, §323), an 'agreement in form of life' (PI, §241): essentially a blind agreement in our shared natural behaviour and human practices. It is this 'consensus of *action*: a consensus of doing the same thing, reacting in the same way' (LFM, §§183–4) – that is at the normative root of our grammar and our concepts. We might say that it is this 'shared sense of the obvious' (Williams, 1999, p. 206) in our form of life that grammatical or normative 'propositions' formulate. So that where our rules of grammar have their root in facts, their anchorage is effected in and through practice, not decision. The objectivity or autonomy of grammatical rules is guaranteed by the blindness with which they are intersubjectively established and followed.

In *On Certainty*, then, Wittgenstein realizes how more far-reaching grammar is than he previously thought: it includes certainties of our world-picture which, when formulated, resemble empirical and contingent propositions:[10]

> I want to say: propositions *of the form of* empirical propositions, and not only propositions of logic, form the foundation of all operating with thoughts (with language). (OC, §401, my emphasis)
>
> If I say '*we assume* that the earth has existed for many years past' (or something similar), then of course it sounds strange that we should *assume* such a thing. But in the entire system of our language-games it belongs to the foundations. The assumption, one might say, forms the basis of action, and therefore, naturally, of thought. (OC, §411, emphasis original)

These 'propositions' that resemble – are 'of the form of' – yet are not in fact empirical and epistemic propositions[11] are 'propositions which we affirm without special testing; propositions, that is, which have a

peculiar *logical* role in the system of our empirical propositions' (OC, §136, my emphasis) – in fact, rules of grammar. They can be:

1. certainties that were *once learned as empirical or epistemic propositions,* but have become so intersubjectively ingrained and fossilized, that they are no longer part of the wealth of empirical or epistemic propositions of a given community (e.g. modern educated adult) but belong to the 'scaffolding' of their thoughts (OC, §211); e.g. 'The earth is round,' 'Trains arrive in train stations,' 'Human beings can go to the moon'
2. certainties that *we may have learned as children, but as rules, not as questionable empirical facts*: 'Babies cannot speak,' 'People die,' 'People sometimes lie,' 'The earth has existed for a long time'
3. certainties that may never have been expressed or taught; these are either *lived certainties*, or certainties that are *assimilated through repeated exposure*: e.g. 'I have a body,' 'There exist people other than myself,' 'The world exists,' 'The earth is a (large) body on whose surface we move,' 'Trees do not gradually change into men and men into trees,' 'If someone's head is cut off, he is dead and will never live again,' 'People usually smile or laugh when they're happy; cry when they're sad or in pain; yell or snap when they're angry,' 'I recognise the people I regularly live with,' 'The majority of people are not mistaken about their names'.[12]

The basic certainties listed in the last two groups can be called 'universal certainties' or 'universal rules of grammar' in that they belong to the scaffolding of thought of any normal human being.[13] They are rules of grammar that are rooted (nonratiocinatively) in 'very general facts of nature' appertaining to 'the natural history of human beings' (PI, §§230, 415). Any empirical enquiry has to take such universal rules of grammar as 'The world exists', 'Human beings live and die' or 'Newborn babies cannot speak' as part of its *logical* or *grammatical* starting points – its grammar. Questioning any of these would not be a legitimate move in a language-game, but a sign of madness or nonsense or misunderstanding of a more innocuous kind.[14] In fact, it would be like questioning *any* rule of grammar, such as '2+2=4'.

But there is a crucial difference between Wittgensteinian universal grammar and that stipulated by nativists. Where Chomsky's universal grammar is in the brain and consists of symbols or structures, Wittgenstein's grammar is really nothing but a way of acting – a logic *in action:*[15]

> Giving grounds, however, justifying the evidence, comes to an end; – but the end is not certain propositions striking us immediately as

> true, i.e. it is not a kind of seeing on our part; it is our acting, which lies at the bottom of the language-game. (OC, §204)

Let me try to explain. Although we can formulate our rules of grammar (as I have been doing here, and as Wittgenstein often does), this formulation or verbalization is always merely heuristic or pedagogic; an expression of a rule of grammar is never an *occurrence* of the mastery of that rule. Our mastery of grammar cannot meaningfully be expressed in the flow of the language-game;[16] it can only *show* itself in what we do and *in* what we say (e.g. my mastery of the grammatical rule 'There exist beings other than myself' shows itself in my speaking to others or of others). If I were to utter the sentence 'There exist beings other than myself', you would all look at me perplexed. Is this supposed to inform you of something you did not already take for granted? Such an utterance is not a move in a language-game, but an expression of part of the *framework* that allows us to make moves in a language-game; to make statements, ask questions, voice disagreements. In *On Certainty,* Wittgenstein gives several examples where stating (one's mastery of) a grammatical rule in non-heuristic situations causes nothing but perplexity; for example: 'If a forester goes into a wood with his men and says, "*This* tree has got to be cut down, and *this* one and *this* one" – what if he then observes "I *know* that that's a tree"?'[17] (OC, §353).

As Wittgenstein writes in *Remarks on the Foundations of Mathematics*, 'The limits of empiricism are not assumptions unguaranteed, or intuitively known to be correct; they are *ways in which we make comparisons and in which we act'* (RFM, II, §21, my emphasis). That these are rules of grammar is clear: 'Everything that's required for *comparing the proposition* with the facts belongs to the grammar' (BT, §38, my emphasis).

I now turn to a philosopher for whom facts play not only a macroscopic, but a microscopic, role in his account of human life and meaning: John Searle. For Searle, all that we are – as physical, mental, social beings – is reducible to brute physical facts, to electrons and neurons.

7.2 Searle: realism reduced to empiricism

In his most recent book, *Making the Social World*, Searle points out the puzzling character of social ontology, the apparent paradox in our understanding of social reality:

> How can we give an account of ourselves, with our peculiar human traits – as mindful, rational, speech-act performing, free-will having,

> social, political human beings – in a world that we know independently consists of mindless, meaningless, physical particles? How can we account for our social and mental existence in a realm of brute physical facts? (2010, p. ix)

My immediate response to this question is that we can do so because the world we live in is not *merely* a realm of brute physical facts; it does not consist *only* of mindless, meaningless, physical particles; for I am not a mindless, meaningless, physical particle, and I am part of what the world consists of. Nor can I, or my thoughts, be reduced to mindless, meaningless, physical particles. And if they can not be thus reduced, then there is no paradox about an account of our social and mental existence in a world of brute physical facts.

In saying this, I am not flouting Searle's first condition for the adequacy of accounts of ourselves in our world: I am not postulating two or three worlds or different ontological realms; but am mindful of giving – as he stipulates – 'an account of how we live in exactly one world, and how all of [the] different phenomena, from quarks and gravitational attraction to cocktail parties and governments, are part of that one world' (2010, p. 3). Before showing that I am not flouting Searle's first condition, let me first mention his argument for biological naturalism, the view that mental phenomena are emergent, higher-level properties of physical or biological systems; they are caused by lower-level neurophysiological processes in the brain and are themselves features of the brain – that is, they are *realized* in the structure of the brain.

Because there is no scientific knowledge of these processes, Searle's argument relies on an analogy. He claims that just as the relation between the molecular structure of a piston and its solidity is one of causation, so conscious states are caused by lower-level neurobiological processes in the brain and are themselves higher-level features of the brain. But I see a problem – as have others before me – in taking the relation between the molecular structure of solidity and solidity to be one of causation; for it seems obvious that the molecular structure of solidity *constitutes* solidity; it does not *cause* it. The molecular configuration of the piston is spatially and temporally co-extensive with its solidity; it does not exist independently of it, and so there is no room here for causation. 'Solidity' is a concept or word we use to refer to what happens when certain molecules attach in a certain way; it is a more economical description of the molecular configuration, not a different-level *phenomenon*. So that giving 'solidity' an ontological status is a category mistake.

To the objection that the higher-order component ought to be eliminated from an account of consciousness, Searle replies that:

> if we did, we would still have the subjective experiences left over. ... We need a word to refer to ontologically subjective phenomena of awareness or sentience. And we would lose that feature of the concept of consciousness if we were to redefine the word in terms of the causes of our experiences. (Searle, 1998, p. 1941)

But I do not see that we would lose 'the subjective feature of the concept of consciousness', what we would lose is Searle's account of it as an emergent higher-level phenomenon. And some of us do not mind losing that account; for, *pace* Searle, his is not the only way we can investigate and account for the subjective feature of consciousness: we can do so by investigating behaviour rather than molecules. And if, as he claims, the point of having the concept of consciousness was to have a word to name subjective experiences, then by all means, let us use the word – only not to refer to an emergent higher-level, ontologically thick phenomenon, but rather to a way of describing 'subjective phenomena of awareness or sentience', as in the sentence: 'I was so weak I thought I'd lose consciousness.'

And what of collective intentionality? Inasmuch as collective intentionality is a type of intentionality, it must, on Searle's view, be mental (that is, caused by and realized in neurobiology). As says Searle: 'If you understand electrons and elections right you will see why some electrons have to participate in elections. No electrons, no elections' (2010, p. 3). So that although institutional facts (the products of collective intentionality; e.g. marriage, academia, the economy) have themselves no physical realization, they still need to bottom out in the entities of physics and chemistry. And what better, or who better, to serve that purpose than human beings? The brute facts, in the case of institutions, are 'actual human beings and the sounds and marks that constitute the linguistic representations' (2010, p. 109) that generate and maintain normative constraints. These constraints are not always explicitly formulated or enforced, and this is where the Background comes in. Collective intentionality, like all intentionality, is possible only against a Background of nonintentional capacities, practices, habits and presuppositions, some of which constitute sets of power relations; So that where the Background has to do with institutions, its norms function as power mechanisms or *standing Directives* wielded directly or counterfactually by human beings (2010, 158–60). For Wittgenstein, too, rule-following

is a matter of the 'quiet agreement' of a community of people, but the importance of the individual in Wittgenstein's communal picture does not reside in her having a brain and thus being a biological generator of *standing Directives* or speech-acts. Rather, her quiet agreement needs no bottoming out – which is not to say that brute physical facts are disregarded, but only that they are not regarded as generative or explanatory of social institutions.

In a sense, of course there is no question that society or social institutions are *caused* by human beings: human beings do bring social institutions about; they cause them to exist. And, of course, human beings have neurobiological structures and processes and, indeed, could not create or cause anything at all, or even think, without these. But it does not follow from the necessity of such structures and processes for thinking that our thoughts are isomorphic or reducible to them:

> there is no copy in either the physiological or the nervous systems which corresponds to a *particular* thought, or a *particular* idea, or memory. (LW I, §504, original emphasis)
>
> Even if we knew that a particular area of the brain is changed by hearing *God Save the King* and that destroying this part of the brain prevents one's remembering the occasion, there is no reason to think that the structure produced in the brain represents *God Save the King* better than *Rule Britannia*. (LPP, §90)

Thinking about quantum physics, talking about my savings or getting married are not dependent on anything molecular other than in the instrumental sense that I am dependent on molecules, but those molecules – just as any other neurobiological conditions for life – are *enabling*, not *determinant*; where an enabling condition is, as Anthonie Meijers puts it: 'one that makes possible a phenomenon, without determining its actual characteristics' (2000, p. 158).[18] The conflation of these – recently expressed by William Ramsey as: 'the characterization of any functional architecture that is causally responsible for the system's performance…as encoding the system's knowledge-base, as implicitly representing the system's know-how' (2007, pp. 3–4) – is what Wittgenstein warned us against. So, yes, Wittgenstein, too, would say: 'no electrons, no elections' and would agree with Searle that 'mental states, such as my present state of consciousness, are caused by a series of neurophysiological events in my brain' (1991, p. 144). But there is a huge leap from that to saying that 'brains cause minds' (1990, p. 29).[19] As Stéphane Chauvier writes: 'That a lesion in a part of her brain prevents

an individual from recognizing certain familiar faces informs us about the neural conditions required for someone to recognize a face, but the recognition of a face is not itself a neuronal process' (2007, p. 46).

A passage from Bennett and Hacker's *Reply to Dennett and Searle* clearly illustrates the confusion in Searle's analogy:

> Professor Searle suggests that the question 'Where do mental events occur?' is no more philosophically puzzling than the question: 'Where do digestive processes occur?' So, he argues, digestive processes occur in the stomach, and consciousness occurs in the brain. This is mistaken. Being conscious, as opposed to unconscious, being conscious of something, as opposed to not noticing it or not attending to it, do not occur *in* the brain at all. Of course, they occur *because of certain events in the brain*, without which a human being would not have regained consciousness or had his attention caught. 'Where did you become conscious of the sound of the clock?' is to be answered by specifying where I was when it caught my attention, just as 'Where did you regain consciousness?' is to be answered by specifying where I was when I came round.
>
> Both digesting and thinking are predicated of animals. But it does not follow that there are no logical differences between them. The stomach can be said to be digesting food, but the brain cannot be said to be thinking. The stomach is the digestive organ, the brain is no more an organ of thought than it is an organ of locomotion. If one opens the stomach, one can see the food being digested there. But if one wants to see thinking going on, one should look at the *Penseur* (or the surgeon operating, or the chess player playing or the debater debating) not at his brain. All his brain can show is what goes on there *while he is thinking*. (Bennett and Hacker, 2007, p. 143)

This brings me to Searle's second adequacy condition, according to which any account of the mental must respect the basic facts of the structure of the universe and show how it is dependent on, and in various ways derives from, those basic facts (2010, p. 4) – the basic facts being those given by physics and chemistry, by the 'atomic theory of matter', by evolutionary biology and the other natural sciences. Well, attention is indeed paid by Wittgenstein to what *he* would call 'basic facts' in his account of our being 'mindful, rational, speech-act performing ... social ... human beings' in a material world, as it were, but Wittgenstein's notion of basic facts (at least, his post-*Tractarian* one) differs from Searle's; it does *not* include anything like the atomic theory of matter, but only facts whose

generality and visibility would fail to satisfy the kind of attention Searle thinks ought to be paid to the sciences in our accounts. But why *should* we go micro? Why should we accept Searle's second adequacy condition to the letter, on his terms? Let us briefly come back to how Wittgenstein relates things like mind, meaning and sociality to world, and see whether that will not do.

As we saw, on Wittgenstein's view, our mindedness, language, rationality and sociality are impacted by 'very general facts of nature' (PI, p. 230) such as the facts that human beings need to eat, sleep, breathe air; that they experience pain and joy and sadness; that apples fall from trees, and cows do not grow on them. A biological fact, such as that the life-span of a human being cannot exceed approximately 125 years, conditions our concept of human life and, more specifically, influences the way we speak about human longevity; so that we cannot sensically speak of a three thousand-year-old man unless it be in archaeological or science-fictional terms. *On Certainty* fleshes out Wittgenstein's claim in *PI* that concepts are conditioned by very general facts of nature, by arguing that some of our normative and grammatical rules – which are part of the background (OC, §94) or, to use another metaphor, part of the hinges on which knowledge turns (OC, §341) – are anchored in (*not grounded in or justified by*, *but anchored in*, or *rooted in*) regularities of nature. Sometimes, then, our grammatical and normative rules *are* made to reflect the regular behaviour of things. As Wittgenstein says, 'The rule we lay down is the one most strongly *suggested* by the facts of experience' (AWL, §84, my emphasis); so that it is through our rules that nature 'makes herself audible'. However, although Wittgenstein is clear that some concepts and grammatical rules are *suggested* or *influenced* by facts, he is also clear about grammar being autonomous; that is, not *answerable* or *accountable* to nature or to facts: 'The essence of logical possibility is what is laid down in language. What is laid down depends on facts, but is not made true or false by them' (AWL, §162).

So that while his acknowledgement of the rootedness of our concepts in the world puts Wittgenstein *outside* the idealist camp, it does not thereby place him in the *realist* camp – and this time, 'realist' in the sense of 'materialist'. There is a difference between saying that certain facts are *favourable to* the formation of certain concepts and saying that our concepts are *dictated* by nature, or (empirically or epistemically) *derived* from nature. The expression '*favourable to*' is meant to show that these facts are not seen as *justifications* but rather, writes Wittgenstein, as 'possible *causes* of the formation of concepts' (PI, p. 230, my emphasis) – where 'cause' is not to be understood as a one-to-one engendering, such

as a flower producing a seed, but as an *influencing* or a *conditioning*. And so, the fact that our concepts are not *founded* on experience does not mean that they are totally divorced from or impervious to it. Our concepts are not *empty,* but how they are informed by the world is not how hard-core realists and empiricists take them to be; they are not rationally or (micro-) causally derived from the world, though they may be *rooted* in it.

So why – when there are other perfectly viable options – must we go micro in any account of ourselves as mindful, social beings? Especially as Searle himself concedes that there is no micro account available:

> a deep understanding of consciousness would require an understanding of how consciousness is caused by, and realized in, brain structures. Right now nobody knows the answers to these questions: how is consciousness caused by brain processes and how is it realized in the brain? (Searle, 2010, p. 26)

Why insist that any account of our mindful, speech-act performing, social selves *must* go all the way down when even science is unable to demonstrate that that is the way to go. If attention is to be paid by philosophy to science, should it not be to scientific *results* rather than to scientific hypotheses that seem nourished by a preconception of how things must be? But this does not stop Searle from stating what is a hypothesis as unquestionable: '... it is just a plain fact that neuronal processes do cause feelings, and we need to try to understand how' (2005, p. 51).

Even if, *per impossibile*, a micro account were available – I say, '*per impossibile*' because it seems to me logically impossible that my mental life can be caused by a configuration of molecules rather than by fully-fledged people, events and things out there in the world: there is no physical or logical room in those molecules for the world. But – again – even if, *per impossibile*, a micro account were available, how would it be more adequate or relevant than a Wittgenstein-type account; how would it be a more *perspicuous presentation* of our human form of life? Water is not relevantly reducible to H_20 other than in very limited contexts, such as chemistry classes and labs: we do not think of, or refer to, as H_20 the stuff we go to the beach to splash in, or the stuff we add to our whisky or admire in Monet paintings or pray to the rain god for. Indeed, those of us who pray to the rain god have probably never even heard of H_20. Moreover, not all H_20 is water: some H_20 we call 'ice'; other H_20 we call 'steam' or 'vapour'.[20] And so, if even water is not – except in very limited

contexts – *relevantly* reducible to its molecular configuration, why should our meanings, feelings and behaviours be? Why should my awareness that I am enjoying the sun or a Chateau Petrus '82 or that I am slowly losing consciousness be things that I would want to attribute to my molecules or thank them for? Or, indeed, find relevance or adequacy in *any* account of myself as a minded and social being who would do so? Anyone in their right mind would accept that we can have no elections without electrons, but what is the explanatory relevance of neuronal processes for our understanding of ourselves as minded beings, or for our understanding of the nature of institutions? As Dan Hutto (personal communication) puts it: 'Something may have to be the case for certain things to happen; but it can be irrelevant to the explanation we need.'

So, to Searle's question about how human beings can create such marvellous features as Declarations and elections, and how they can maintain these in existence once created, I would answer – as he also partly does – through language and practices; but then *I* would go no further. For to go further is to take part in the shadow play of reductionism; as Raimond Gaita reminds us, the most fundamental aspect of Wittgenstein's legacy is that we cannot purify our concepts of their embeddedness in human life without being left with only a shadow play of the grammar of serious judgment (1990, xii). That shadow play is reflected, it seems to me, in the dance of neurons and electrons that make up Searle's ultimately reduced, and therefore dehumanized, world.

Notes

I am grateful to Dan Hutto and Britt Harrison for valuable comments on the initial draft of this chapter, as well as to Gabriele Mras and participants in the Ludwig Wittgenstein Workshop-Series on *Wittgenstein on Concept-Formation and the Limits of the Empirical* held at the WU Vienna University of Economics in December 2011 for stimulating discussion.

1. Empiricism, as Wittgenstein uses it here, in contrast to realism, is the view that experience is rationally based on evidence.
2. That is: philosophical elucidation may also be based on thought experiments.
3. PI Part II (1946–1949) is roughly contemporaneous with RPP I (1946–1947).
4. Of course, there are cases where a rule of grammar may have its historical root in an empirical discovery – example the realization that men have something to do with the reproductive process was at first of an empirical nature – however, it is not this that is of interest to philosophers, but rather the logical link which developed (nonepistemically and nonempirically) from this. More on this later in the chapter. Also, the adoption of some of our rules of grammar may have a pragmatic justification (see note 10 below).

5. Note that in *On Certainty*, 'ground' is usually synonymous with 'reason'.
6. Basic certainties all function as grammatical rules, but not all grammatical rules are basic certainties: one of the features of basic certainties is that they manifest themselves as a thoughtless, flawless know-how, whereas we do not have an automatic or thoughtless grasp of many mathematical and grammatical rules; some require thought, calculation or recollection (e.g. '235 + 532 = 767'; 'a funambulist is a tightrope walker'). The reason I am using grammar and certainty interchangeably here is that the focus of this discussion is on the correspondence between grammatical rules and 'very general facts of nature', such as are always before our eyes, and about which there can be no hesitation or doubt, and therefore grammatical rules here (e.g. 'Human beings are normally susceptible to pain'; 'There is psychological as well as physical pain'; 'Human beings need sleep, nourishment; normally sleep and think') are all also basic certainties. For more elaborate discussion, see Moyal-Sharrock (2007, pp. 118–9).
7. 'The *truth* of certain empirical propositions belongs to our frame of reference' (OC, §83, emphasis original). Notice that *truth* is here italicized, making it clear that it is not truth at all, but what Moore mistakes for truth, that can be attributed to our basic certainties: 'If the true is what is grounded, then the ground is not true, not yet false' (OC, §205); this, further confirming that these apparent empirical propositions are not empirical propositions at all, but only 'have the form of' empirical propositions' (OC, §96).
8. For, remember that the function Wittgenstein accords a rule of grammar is not *only* the narrow one of instructing us in the use of individual words (e.g. 'A rod has (what we call) a length' or 'This is a hand'); but also, more broadly, that of expressing the *conditions* for making sense (PG, p. 88). Such 'conditions' include anything that is 'a preparation for a description' (MWL, §72) – in other words, anything that *underpins* a description. It is because he realizes that our basic certainties *underpin* our descriptions – indeed, our making sense altogether – that Wittgenstein concludes they are expressions of *norms* of description or rules of grammar, and not empirical or epistemic propositions (for 'how can we describe the foundation of our language by means of empirical propositions?' (RFM, §236).
9. Except justification of a pragmatic type: 'The essence of logical possibility is what is laid down in language. What is laid down depends on facts, but is not made true or false by them. What *justifies* a symbolism is its usefulness' (AWL, §162, my emphasis); 'What you are saying is not an experiential proposition at all, though it sounds like one; it is a *rule*. That rule is made important and justified by reality – by a lot of most important things.' If you say, 'Some reality corresponds to the mathematical proposition that $21 \times 14 = 294$', then I would say, Yes, reality, in the sense of experiential (empirical) reality *does* correspond to this. For example, the central reality that we have methods of representing this so that it can all be seen at a glance (LFM, §246, emphasis original). But although it may be an empirical fact that, say, 'men calculate like this! ... that does not make the propositions used in calculating into empirical propositions' (RFM, §381); though the technique is a fact of natural history, its rules do not play the part of propositions of natural history (RFM, §379).

10. These apparent *empirical* propositions revealed by the 'third Wittgenstein' (that is, to Wittgenstein's work from 1946 on) to be rules of grammar extend the scope of the apparent *super-empirical* or *metaphysical* propositions shown up by the second Wittgenstein to be rules of grammar. For a more elaborate discussion of the difference, see Moyal-Sharrock (2004, 2007).
11. 'That is, we are interested in the fact that about certain empirical propositions no doubt can exist if making judgments is to be possible at all. Or again: *I am inclined to believe that not everything that has the form of an empirical proposition is one*.' (OC, §308, my emphasis)
12. Most of these examples are drawn from *On Certainty*.
13. This makes it clear that I am not using the term 'universal' as: 'applicable to all possible worlds'.
14. Such as that which results from not properly hearing what was said or from not properly mastering the language in which it was said.
15. For more on this, see Moyal-Sharrock (2003)
16. 'So if I say to someone "I *know* that that's a tree" ... a philosopher could only use this statement to show that this form of speech is actually used. But if his use of it is not to be merely an observation about English grammar, he must give the circumstances in which this expression functions' (OC, §433). This, of course, does not preclude our formulating rules of grammar in heuristic or pedagogical situations. For a more elaborate discussion of the (technical) ineffability of grammatical rules in the flow of the language-game – their being 'removed from the traffic' (OC, §210) of ordinary discourse, see Moyal-Sharrock (2007, 65ff, 94ff).
17. See also pp. 460–9. Here is OC, §468: 'Someone says irrelevantly "That's a tree". He might say this sentence because he remembers having heard it in a similar situation; or he was suddenly struck by the tree's beauty and the sentence was an exclamation; or he was pronouncing the sentence to himself as a grammatical example; etc., And now I ask him "How did you mean that?" and he replies "It was a piece of information directed at you". Shouldn't I be at liberty to assume that he doesn't know what he is saying, if he is insane enough to want to give me this information?'
18. See also Moyal-Sharrock (2000, p. 356), which rebukes physicalism for thinking of the brain as '[n]ot simply a mechanical enabler, [but] the generator of our wills, desires, intentions and actions.'
19. Or that 'brains attach meaning to minds' (Moyal-Sharrock, 2000, p. 26). The article's title reads: 'Is the Brain's Mind a Computer Program? No. A program merely manipulates symbols, whereas a brain attaches meaning to them'. How is that done? The answer we get is never much more than mere reiteration: '... the mind imposes intentionality on sounds and marks, thereby conferring meanings upon them and, in so doing, relating them to reality.'(Searle, 1999, p. 139) For Searle: 'actual human mental phenomena [are] dependent on actual physical-chemical properties of actual human brains' (1990, p. 29).
20. I am inspired, here, by Avrum Stroll's (1998, pp. 37–73) excellent 'Reflections on Water' in his *Sketches of Landscape: Philosophy by Example*.

References

M. R. Bennett and P. M. S. Hacker (2007) *Reply to Professor Dennett and Professor Searle* in *Neuroscience & Philosophy: Brain, Mind and Language* (New York: Columbia University Press).

S. Chauvier (2007) 'Wittgensteinian Grammar and Philosophy of Mind' in D. Moyal-Sharrock (ed.) *Perspicuous Presentations: Essays on Wittgenstein's Philosophical Psychology* (Basingstoke, UK: Palgrave Macmillan, 2007), 28–49.

R. Gaita (1990) 'Introduction' in *Value and Understanding: Essays for Peter Winch* (New York: Routledge), ix–xiii.

R. Harré, and E. H. Madden (1975) *Causal Powers: A Theory of Natural Necessity* (New Jersey: Rowman and Littlefield).

A. W. M. Meijers (2000) 'Mental Causation and Searle's Impossible Conception of Unconscious Intentionality', *International Journal of Philosophical Studies*, 8(2), 155–70.

D. Moyal-Sharrock (2000) '"Words as Deeds": Wittgenstein's "Spontaneous Utterances" and the Dissolution of the Explanatory Gap', *Philosophical Psychology*, 13(3), 355–72.

—— (2003) 'Logic in Action: Wittgenstein's *Logical Pragmatism* and the Impotence of Scepticism', *Philosophical Investigations*, 26(2), 125–48.

—— (2004) 'On Certainty and the Grammaticalisation of Experience' in D. Moyal-Sharrock (ed.) *The Third Wittgenstein: The Post-Investigations Works* (Aldershot, UK: Ashgate, 2004), 43–62.

—— (2007) *Understanding Wittgenstein's On Certainty* (Basingstoke: Palgrave Macmillan).

W. Ramsey (2007) *Representation Reconsidered* (Cambridge: Cambridge University Press).

Searle, John (1999) *Mind, Language and Society: Doing Philosophy in the Real World* London: Weidenfeld and Nicolson).

J. R. Searle (1990) 'Is the Brain's Mind a Computer Program?' *Scientific American,* 262, 26–31.

—— (1991) 'Response to Freeman/Skard' in E. Lepore and R. Van Gulick (eds) *John Searle & His Critics* (Cambridge: Basil Blackwell), 141–6.

—— (1998) 'How to Study Consciousness Scientifically', *Phil. Trans. R. Soc. Lond.*, (353), 1935–1942.

—— (2005) 'Reply to Stevan Harnard "What Is Consciousness?"' *The New York Review of Books*, 52(11), (online).

—— (2010) *Making the Social World: The Structure of Human Civilization* (Oxford: Oxford University Press).

—— (2011) 'Wittgenstein and the Background', *American Philosophical Quarterly*, 48(2), 119–28.

Stroll, A. (1998) *Sketches of Landscape: Philosophy by Example* (Cambridge, MA: MIT Press).

Williams, Meredith (1999) *Wittgenstein, Mind and Meaning: Toward a Social Conception of Mind* (London: Routledge).

8

Can a Robot Smile? Wittgenstein on Facial Expression

Diane Proudfoot

8.1 Introduction

Recent work in social robotics, which is aimed both at creating an artificial intelligence and providing a test-bed for psychological theories of human social development, involves building robots that can learn from 'face-to-face' interaction with human beings – as human infants do. The building-blocks of this interaction include the robot's 'expressive' behaviours, for example, facial-expression and head-and-neck gesture. There is here an ideal opportunity to apply Wittgensteinian conceptual analysis to current theoretical and empirical work in the sciences. Wittgenstein's philosophical psychology is sympathetic to embodied and situated Artificial Intelligence (see Proudfoot, 2002, 2004b), and his discussion of facial-expression is remarkably modern. In this chapter, I explore his approach to facial-expression, using smiling as a representative example, and apply it to the canonical interactive face robot, Cynthia Breazeal's Kismet (see e.g. Breazeal, 2009, 2002). I assess the claim that Kismet has expressive behaviours, with the aim of generating philosophical insights for AI.

Section 8.2 describes aspects of recent work in the sub-field of social robotics that I call *face AI*. Sections 8.3 and 8.4 analyse and elaborate on Wittgenstein's remarks concerning facial-expression and other expressive behaviour, in order to answer the question *Can a face robot smile?* I unpack the significance of Wittgenstein's remark, 'A smiling mouth *smiles* only in a human face' (PI, §583). Smiling is more than the production of a certain physical configuration or behaviour. It is a complex conventional gesture. A facial display is a *smile* only if it has a certain meaning – the meaning that distinguishes a smile from a human grimace or facial tic, and from a chimpanzee's bared-teeth display.

According to Wittgenstein, a display acquires this meaning only if it is part of a 'normal play' of mobile human expressions and is appropriately embedded in the 'bustle of life'. To be *smiles*, the displays of face robots must satisfy this requirement.

Section 8.5 applies Wittgenstein's remarks on expressive behaviour to human infants, and compares the infant with the 'child-machine' – a machine that is to learn human-like behaviours autonomously, aided by human scaffolding. Although neither the infant's nor the child-machine's 'smile' suffices for *smiling*, the infant's display is closer to satisfying the conceptual requirements on expressive behaviour. Section 8.6 analyses Wittgenstein's remarks on the recognition of facial expressions and on (what is now called) mind-blindness. This yields a philosophical characterization of an 'expressive' face robot as a *mind-blind smiling-machine*.

A Wittgensteinian perspective on the claims of ('strong') face AI makes three contributions. First, it generates new conceptual claims about expressive behaviour and facial-expression recognition, which are also the bases of testable empirical hypotheses. Second, it provides a conceptual framework that is valuable in tackling (what I call) the forensic problem of anthropomorphism for AI (Proudfoot, 2011). Our tendency to anthropomorphism stands in the way of AI's grand goal of building a physical machine that thinks; any supposed demonstration of artificial intelligence must exclude the possibility that we are merely anthropomorphizing an unintelligent machine. Wittgenstein's distinction between different senses of 'smiling' ('frowning', and so on) helps to avoid the misplaced anthropomorphism of social robots. Third, Wittgenstein's ideas point to crucial differences between the human child and the *child-machine*, a fundamental concept in machine intelligence since Turing.

8.2 Face AI

Despite the confidence of some early researchers, 'Good Old-Fashioned AI' failed to produce a human-like artificial intelligence, and ran into seemingly intractable problems – for example, the difficulty of building a system capable of 'common-sense' knowledge. This led in the 1980s and 1990s to a greater emphasis upon embodied and socially situated AI (see Brooks, 1999). The grand aspiration of building what Turing called a child-machine (Turing, 1950) includes implementing on a robot 'theory of mind' abilities, such as face and agent detection, joint visual attention, self-recognition, and success in false belief tasks (e.g. Scassellati, 2000, 2002; Gold and Scassellati, 2007; Breazeal, Gray,

and Berlin, 2009). Other abilities of pre-verbal infants to be implemented include the detection of prosody in a human voice and of intent in visual motion (Scassellati et al., 2006; Kim and Scassellati, 2007; Breazeal and Aryananda, 2002), and arm reaching and imperative pointing (e.g. Sun and Scassellati, 2005; Scassellati, 2000). Social roboticists aim to use results from cognitive neuroscience and developmental psychology to implement these basic behaviours on a robot, and then to bootstrap more complex behaviour using strategies analogous to those of human infant-carer (or sibling-sibling) interaction (see e.g. Breazeal, 2009). The goal is a 'socially intelligent' robot. (For an overview of the principles underlying this work in human–robot interaction, see Dautenhahn, 2007.) The robots involved include Cog (e.g. Brooks et al., 1999), Kismet (Breazeal, 2002), Leonardo (Breazeal, 2006, 2009; Breazeal et al., 2005), Nexi (Breazeal, 2009), Bandit (e.g. Wade et al., 2012), Nico (Sun and Scassellati, 2005; Scassellati et al., 2006), Infanoid (Kozima, Nakagawa, and Yano, 2005) and Mertz (Aryananda, 2006).

Researchers in social robotics also have narrower aims. These include: testing hypotheses about human social and psychological behaviour and development (and about cognitive developmental disorders such as autism), by investigating human-robot interaction in standardized conditions (e.g. Scassellati, 2005); building artificial systems that are capable of quickly learning new behaviours from (and responding accurately and sensitively to) humans, by means of normal social cues (e.g. Breazeal, 2006); and producing service, entertainment and therapeutic robots, with which humans with no specialized training can interact intuitively (Scassellati, 2000, 2005; Tapus, Matarić, and Scassellati, 2007; Fasola and Matarić, 2012; Wade et al., 2012; Feil-Seifer and Matarić, 2011; Giullian et al., 2010). The study of human-robot interaction and of 'expressive' anthropomorphic robots has expanded hugely in recent years (for a taxonomy of social robots and of design methodologies, see Fong, Nourbakhsh, and Dautenhahn, 2003).

Turing suggested in 1950 that one approach to machine intelligence would be to provide a machine with 'the best sense organs that money can buy', and then 'teach it to understand and speak English' (1950, p. 460). Modern roboticists aim to build some 'sense organs' (restricted vision and hearing, and recently olfaction), but typically not to teach their robots English. (On natural-language-competent robots, see Shapiro, 2006.) Instead, the goal is to give the machines developmentally-earlier communicative abilities, including facial-expression and

Figure 8.1 Janet showing happiness. Reprinted with permission of Chyi-Yeu Lin

head-and-neck gesture. This sub-field of AI – constructing robots with a 'face' and 'face robots' (machines consisting only of a 'head' and 'face'), and giving them 'facial expressions' – will here be called *face AI*.

A face robot may be quasi-realistic, constructed to resemble a human face and head – for example, Fumio Hara and Hiroshi Kobayashi's series of face robots (Hara 2004; Hara and Kobayashi, 1997) and Chyi-Yeu Lin's face robot singer (Lin et al., 2011) (Figure 8.1)[1] – or it may be a caricature. (On how people perceive robots with different appearances, see e.g. Lee, Lau, and Hong, 2011.)

The canonical expressive interactive face robot is Cynthia Breazeal's (now retired) Kismet; this machine has limited movement, visual and auditory systems, and prosodic 'vocalization' (Breazeal, 2002, 2003c) (Figure 8.2). I shall take Kismet as a paradigm example of an expressive face robot. It is designed to provide 'emotional feedback' and to convey 'intentionality' through facial expressions and behaviour (Breazeal, 2001; Breazeal and Fitzpatrick, 2000). According to Breazeal, 'The robot is able to show expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise' (Breazeal and Scassellati, 2000). Her aim is to 'build robots to engage in meaningful social exchanges with humans,' and 'the key task for [Kismet] is to apply various communication skills acquired during social exchanges' (Breazeal, 1998a). Kismet,

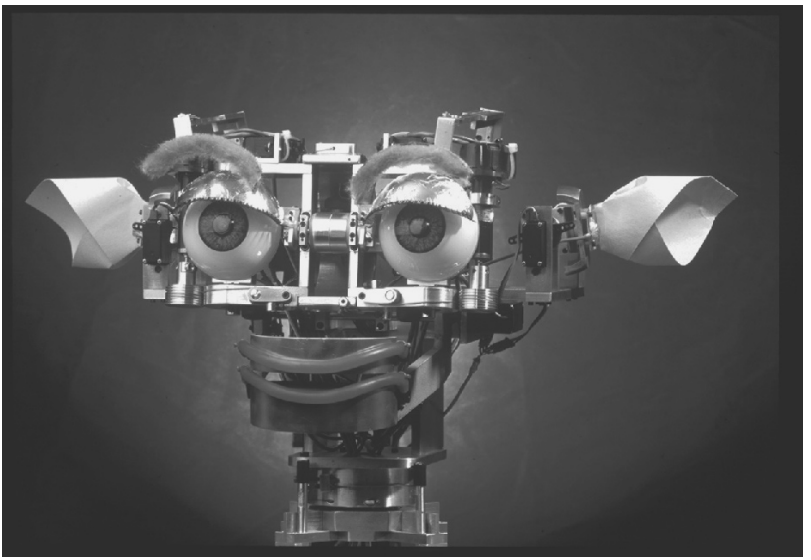


Figure 8.2 Kismet showing happiness. © Peter Menzel/menzelphoto.com. Reprinted with permission

Source: From the book *Robo Sapiens: Evolution of a New Species*, by Peter Menzel and Faith D'Aluisio; copyright Menzel Photo Archives.

Breazeal claims, 'doesn't engage in adult-level discourse, but its face serves many of these functions at a simpler, pre-linguistic level' (2002, p. 157).

A social robot like Kismet is designed to be a 'believable creature' (believability is especially important in socially assistive robotics – see Tapus, Matarić, and Scassellati, 2007).[2] Kismet has caricature child-like 'features', to prompt infant-carer behaviour in the human observer (see Section 8.5). The robot's face responds 'in a timely manner to the person who engages it as well as to other events in the environment', Breazeal says (2002, p. 157). For example, if Kismet 'finds something like a colourful block, it will look at it with a look of happiness because it has found what it wanted. If you play with the robot nicely with the toys, it smiles'.[3] Kismet's 'expressions' are intended to be readable; an untrained human observer should be able to predict and explain the robot's behaviour, by making inferences about the 'emotions' ('goals' and 'needs') manifested in the robot's 'facial' displays (Breazeal, 1999). The aim is that the observer respond to these displays as if to ordinary human social signals, with the result that the displays constrain

the human's behaviour in a manner that suits the robot (e.g. Breazeal, 2003a, 2003b). Human observers typically respond to Kismet in exactly this way. For example, one observer 'intentionally put his face very close to the robot's face ... The robot withdrew while displaying full annoyance in both face and voice. The subject immediately pushed backwards, rolling the chair across the floor to put about an additional three feet between himself and the robot, and promptly apologized to the robot' (Breazeal and Fitzpatrick, 2000). Some observers even mirror Kismet's behaviour, as they might another human being's behaviour (Breazeal, 2001).

Can a robot *have* emotions? This is currently a much-discussed question in AI (see e.g. Arbib and Fellous, 2004; Adolphs, 2005).[4] Breazeal says that Kismet's 'emotions' are 'quite different' from emotions in humans (Breazeal, 1998a; see also Breazeal, 1999). Kismet's 'emotions' are designed to be 'rough analogs' of human emotions (Breazeal, 1998a). Three similarities between robot and human are implicit in the literature on Kismet. The robot's *system architecture* is influenced by theories in developmental psychology and emotion theory: its 'motivation system' has 'drive' and 'emotion' subsystems that are inspired by work in human ethology. This architecture guides Kismet's interaction with humans – the robot's 'expressions' are designed to evoke responses in humans that conform to its 'goals' and 'drives'. The robot's behaviour is *responsive to the social environment* in ways similar to human social behaviour. For example, given the appropriate states of the 'drive' system, if a human observer 'engages the robot in face-to-face contact ... the robot displays *happiness*'; if the slinky toy is removed, 'an expression of *sadness* returns to the robot's face' (Breazeal and Scassellati, 2000).[5] The configurations of the robot's 'facial features' and 'body postures' also copy a human's *expressive behaviours*; 'The robot's facial features move analogously to how humans adjust their facial features to express different emotions', Breazeal claims (1998a).

Typically Breazeal uses scare-quotes or other notational devices when describing the states of Kismet's 'emotion' system. She remarks, for example, that the robot 'responds with an expression of "happiness"', and that 'an expression of "anger" is blended with the intensifying look of "fear"' (Breazeal and Scassellati, 2000, pp. 21–2). In contrast, Breazeal and other researchers ascribe expressive behaviours to the robot literally and without qualification – thus Kismet 'smiles' (Breazeal, 2000). Kismet is also said, without scare-quotes, to have a 'smile on [its] face', a 'sorrowful expression', a 'fearful expression', a 'happy and interested expression', a 'contented smile', a 'big grin' and

a 'frown' (Breazeal, 2001, pp. 584–8; Breazeal and Scassellati, 2001; Breazeal 2000). However, *does* the robot smile, or is this misplaced anthropomorphism?

8.3 Anatomical versus embedded smiling

According to Wittgenstein, a person who is shown a drawing such as this, and asked to describe what he or she sees would immediately say, 'A face' (PI II, p. 204). This is simply the 'best description I can give of what was shewn me for a moment' (PI II, p. 204). Moreover, anyone shown the drawing 'will be able straight away to reply to such questions as, "Is it male or female?", "Smiling or sad?", etc.' (BBB, p. 163). But we are not under 'the delusion of seeing a "real" face', Wittgenstein said (BBB, p. 163). So, how can a *drawing* be smiling or sad?

Wittgenstein used the idea of a '*purely* visual concept' – this is a concept that is used 'purely to describe the structure of what is perceived' (LW I, §§736, 739). Applied to a drawing like the one above, the word 'sad', he said, 'characterizes the grouping of lines in a circle' (PI II, p. 209). It is in this sense too that a drawing can be 'smiling'. The purely visual concept of smiling refers simply to, or is based on, the anatomy of the human face; the 'grouping of lines in a circle' of a 'smiling' drawing corresponds to the facial configuration of the smiling human being. I shall call smiling in the purely visual sense *anatomical smiling* (and likewise for frowning, crying, and other emoting). In this sense, a human smiles when only going through the motions of smiling – for example, in response to the orthodontist's instruction to 'smile into the mirror'. In this sense too an emoticon smiles – and Kismet smiles (and also has an angry or sorrowful expression).

According to Wittgenstein, the word 'sad' can in addition be used in a sense that has '*more* than purely visual reference', where it has 'a different, though related, meaning' (PI II, p. 209). Applied to a human being, he said, the concept of sadness is not a purely visual concept – as

is shown by the fact that 'sadness I can also hear in his *voice* as much as I can see it in his face' (LW II, §755). Other concepts that might be thought purely sensory likewise contain more. For example, *wailing*: we can use the word 'wailing' to 'describe what is purely acoustical', Wittgenstein said. 'But the truth of the matter is: "Wailing" is not a purely acoustical concept' (LW I, §748). Used of a human being, perhaps the word indicates distress as much as sheer sound.

The word 'smiling' too has a sense that is not purely visual – Wittgenstein remarked that we can speak with 'a *smiling* tone of voice' (LW I, §39, his emphasis). In this more-than-purely-visual sense, what does 'smiling' mean? Wittgenstein said, 'A friendly mouth, friendly eyes, the wagging of a dog's tail are primary symbols of friendliness: they are parts of the phenomena that are called friendliness' (PG, p. 28). According to Wittgenstein, what a behaviour symbolizes is determined by the 'form of life' in which the behaviour occurs; he said, for example, that 'hope, belief, etc. [are] embedded in human life, in all of the situations and reactions which constitute human life' (RPP II, §16). For this reason, I shall call smiling in the more-than-purely-visual (or symbolic) sense *embedded smiling* (see further Section 8.4).

Wittgenstein called the drawing above a 'picture-face' (i.e. a picture of a face). He said, 'In some respects I stand towards [the picture-face] as I do towards a human face. I can study its expression, can react to it as to the expression of the human face. A child can talk to picture-men or picture-animals, can treat them as it treats dolls' (PI II, p. 194). This is to anthropomorphize a mere drawing, but only in play. The situation is different with Kismet, however. The descriptions of the robot in the AI literature do not distinguish the anatomical from the embedded sense of 'smile' ('frown', and so on). Developmental robotics may, in consequence, appear to have achieved the goal of building a robot with early communicative abilities – a machine, that is, that smiles and frowns in the embedded sense – *just because* Kismet 'smiles' in the anatomical sense. Such anthropomorphizing is not harmless.

On the other hand, perhaps Kismet's 'emotion' system and situated behaviour *are* sufficient for the robot's motor acts and 'facial' displays to be genuine expressive behaviour. What does embedded smiling require, and does Kismet smile in this sense?

8.4 The Wittgensteinian challenge to 'strong' face AI

Applied to face robots, the distinction between anatomical smiling and embedded smiling evokes John Searle's famous distinction between

'weak AI' and 'strong AI' (Searle, 1980). As Searle uses these terms, weak AI is concerned to *simulate* human cognition in a computer; the overall aim is to build machines that, as Marvin Minsky said, 'do things that would require intelligence if done by men' (Minsky, 1968, p. v). Strong AI, in contrast, is concerned to *duplicate* human cognition in a computer; the overarching goal is to build, as John Haugeland said, 'the genuine article: *machines with minds,* in the full and literal sense' (Haugeland, 1985, p. 2). The objective of (what I shall call) *weak face AI* is to build machines that smile in the anatomical sense. The objective of *strong face AI* is to build machines that smile in the embedded sense.

For Searle, duplicating human cognition in a computer requires that the machine have the 'awareness' or 'mental life' that humans possess by virtue of their biology; in his view, '[o]nly a being that could have conscious intentional states could have intentional states at all' (1980, pp. 454, 452; 1992, p. 132). Wittgenstein's approach to cognition is very different; in place of (what he called)[6] 'gaseous' conscious states, he emphasized the importance of behaviour, history and environment (see Proudfoot, 2004a, 2004b; on similarities and differences between Wittgenstein's remarks and Searle's Chinese room argument, see Proudfoot, 2002). His account of facial-expression provides (at least part of) the conceptual requirements for strong face AI.

8.4.1 Behaviour's subtle shades

Wittgenstein said, '"Facial expression" exists only within a play of the features' (LW I, §766). The conceptual distinction between *anatomy* and *expression* requires that facial expressions (and gestures and postures) *vary*:

> Suppose someone had always seen faces with only *one* expression, say a smile. And now, for the first time, he sees a face changing its expression. Couldn't we say here that he hadn't noticed a facial expression until now? Not until the change took place was the expression meaningful; earlier it was simply part of the anatomy of the face. (RPP II, §356)

'Smiling' is 'our name for an expression in a normal play of expressions', Wittgenstein said (Z, §527). *Smiling* is located in a particular space of facial expressions, just as C# belongs in a particular musical scale.

According to Wittgenstein, a smile must also be *mobile*:

> [I]f one were trying to imagine a facial expression not susceptible of gradual and subtle alterations; but which had, say, just five positions;

> when it changed it would snap straight from one to another. Would this fixed smile really be a smile? And why not? – I might not be able to react as I do to a smile. Maybe it would not make me smile myself. (RPP II, §614)

'Variability and irregularity are essential to a friendly expression. Irregularity is part of its physiognomy', he said (RPP II, §615). (Interestingly, smiles (in adults) are typically irregular, with greater intensity on the left side of the face – see Nagy, 2012.) Other expressions too must be mobile – would we, Wittgenstein asked, say that a person is in *pain* 'if he always produced exactly the same suffering expression?' (LW II, p. 67). Likewise, grief – 'The facial features characteristic of grief ... are not more meaningful than their mobility' (RPP II, §627). This is an example of the 'importance we attach to the subtle shades of behaviour', Wittgenstein said (RPP II, §616). It is in this sense that a 'smiling mouth *smiles* only in a human face'.

The claim that variability and mobility are essential to facial expressions is an empirical hypothesis about how human beings actually react to faces, which can be tested. But it is also a conceptual claim that is implied by Wittgenstein's broader philosophical psychology. In his view, the 'human body is the best picture of the human soul' (PI II, p. 178). The 'soul' that can be seen in a face is just an *aspect* of that face – 'And if the play of *expression* develops, then indeed I can say that a soul, something *inner*, is developing', Wittgenstein said (LW I, §947; see further Section 8.6). This approach to the mind has the consequence that, without the normal play of mobile expressions, 'behaviour would be to us as something completely different' (RPP II, §627). Without 'the subtle shades' of behaviour, there is no *soul*, no *inner states* – and no *smiling* in anything but the anatomical sense. There is only 'machine' behaviour.[7]

Although Wittgenstein rejected dualism, his philosophical psychology generates a tough requirement for Kismet's creators. If the robot is to be said to smile in the embedded sense, it must have a normal range and play of human facial expressions and a mobile 'smile'. According to Breazeal, Kismet's 'facial features move analogously to how humans adjust their facial features to express different emotions' (see Section 8.2). Certainly, the robot's 'features' do move between a number of 'expressions', and the transition from one expression (or posture) to another is smooth and in real time. However, like Lin's face robot singer and other face robots such as ROMAN (Berns and Hirth, 2006), Kismet hardly has a 'normal play' of expressions, and its 'expressions' are *not* variable or irregular – it behaves

'mechanically' (see Section 8.5). From a Wittgensteinian perspective, Kismet does *not* smile (in the embedded sense).

But is this merely a technological problem? Researchers in human-robot interaction whose aim is to design believable anthropomorphic robots can be seen as trying to capture the 'subtle shades' of human behaviour (see e.g. Lee, Lau, and Hong, 2011). Let us imagine, then, that advances in engineering make possible a descendant of Kismet that produces (anatomical) 'smiles' as part of a normal play of mobile expressions.[8] Would this suffice for strong face AI?

8.4.2 When situated ain't situated

Smiling is a sophisticated behaviour, with multiple meanings. As Wittgenstein pointed out, I can even say 'I'm afraid' with a smile (LW I, §21). (On the 'smile of pain', see Kunz, Prkachin, and Lautenbacher, 2009.) Facial expressions work as 'signals', he said, and he sometimes spoke of 'smile-signs' and 'frown-signs' (LPP, p. 283). AI researchers aiming to build robots that actually have emotions focus on *feelings* (e.g. Adolphs, 2005). However, the difference between an anatomical 'smile' and an embedded *smile* is not that the latter is accompanied (or caused) by a feeling of happiness (or some other emotion), since such psychological accompaniments are insufficient to give *meaning* to facial displays.[9] In Wittgenstein's view, whatever the biological origins and underlying neural mechanisms of smiling, human beings are *trained* to smile, and the significance of smiling derives from this training (RPP I, §131).

What makes a 'smile' a *smile* is its role in a 'language game' or 'form of life'. This is an example of a general phenomenon – according to Wittgenstein, one can move chess pieces (in accordance with the rules of chess) yet not be *playing chess*, or make 'Chinese noises' yet not be *speaking Chinese* (LPP, p. 55), or exhibit some of the expressive behaviour associated with grief yet not be *grieving*:

> 'Grief' describes a pattern which recurs, with different variations, in the weave of our life. If a man's bodily expression of sorrow and of joy alternated, say with the ticking of a clock, here we should not have the characteristic formation of the pattern of sorrow or of the pattern of joy. (PI II, p. 174)

In addition, the same 'smile' can be embedded in different behaviour and so acquire different meanings; a 'grin of friendship and grin of rage may be visually similar, but the consequences are different',

Wittgenstein said (LPP, p. 39). A 'smile' takes its meaning from its context:

> I see a picture which represents a smiling face. What do I do if I take the smile now as a kind one, now as malicious? Don't I often imagine it with a spatial and temporal context which is one either of kindness or malice? Thus I might supply the picture with the fancy that the smiler was smiling down on a child at play, or again on the suffering of an enemy. (PI, §539)

For Wittgenstein, expressive behaviour includes not only 'the play of facial expression and the gestures' but also 'the surrounding, so to speak the occasion of this expression' (RPP I, §129). We identify an action 'according to its background within human life'; this background – 'the whole hurly-burly' of human actions – 'determines our judgment, our concepts, and our reactions' (RPP II, §§624–9). Thus we see a facial display as a *smile* or hear a vocalization as a cry of *pain* by locating it within what Wittgenstein called 'the bustle of life' (RPP II, §625). The concept of smiling, like our other concepts, 'points to something within *this* bustle' (RPP II, §625). A 'smiling mouth' *smiles* only when located in this bustle.

Social roboticists recognize that the meaning of a facial-expression is context-dependent (see e.g. Hara and Kobayashi, 1997, p. 586). According to Breazeal, Kismet's 'facial expression' *is* a 'signal' to the human observer – the robot's 'sorrowful expression', for example, is 'intended to elicit attentive acts from the human' (2001, pp. 584–5). And the robot's behaviour *is* socially situated; for example, Kismet produces 'smiles' when, after a lack of stimuli, a human being comes into sight – just as a human might do. Nevertheless, from a Wittgensteinian perspective, Kismet's 'expressions' have merely minimal 'surroundings' (to use Wittgenstein's term). This can be seen simply by asking: What *sort* of smile is the robot's 'smile' – friendly, kindly, or malicious? Kismet's 'smile' is none of these, since the robot does not (in fact) have any of the behaviours that are associated with friendliness, or kindness, or malice, and that turn a 'smile' into a *smile*. Likewise for Kismet's 'sad' expression – is the robot grieving, depressed, or sulking? The answer is again: none of these. The consequence is that, contrary to the roboticists' claims, the robot does not *smile*, or have a *sorrowful* expression.

This is (what I shall call) the *embedding problem* for strong face AI. Unless a robot's behaviour is appropriately embedded, to say that the robot *smiles* is misplaced anthropomorphism. We could introduce a

technical concept – *smiling** – for which embedding in the 'bustle of life' is unnecessary; but this would not suffice for strong face AI. Even if a future Kismet were to have much greater variability and irregularity of 'facial' displays, strong face AI requires that the robot's displays be smiling in 'the full and literal sense', not merely smiling*.

8.5 Baby smiles

People interact with Kismet in some respects as if the robot were an infant, for example speaking to it with the overdone prosody typical of 'motherese' (Breazeal and Aryananda, 2002). Social robots such as Kismet are modelled on human infants. The goal of developmental roboticists is that their robots, from an initial state of mere 'motor babbling' and aided by scaffolding from humans, learn early human social behaviours in a way that is biologically plausible. The researchers' vision is of 'a machine that can learn incrementally, directly from human observers, in the same ways that humans learn from each other' (Scassellati et al., 2006, p. 41; see Shic and Scassellati, 2007). It is in this sense that Kismet is a 'child-machine'.

According to Breazeal, Kismet is constructed 'to receive and send human-like social cues to a caregiver, who can ... shape its experiences as a parent would for a child' (Breazeal and Fitzpatrick, 2000, p. 2). A human infant, Breazeal says, naturally 'displays a wide assortment of emotive cues during early face to face exchanges with his mother such as coos, smiles, waves, and kicks'. The mother interprets this behaviour 'as meaningful responses to her mothering and as indications of his internal state'. The infant 'does not know the significance' of his behaviour for his mother, but learns the actions that produce specific responses from her; over time mother and infant 'converge on specific meanings' for these actions.[10] Kismet is designed to generate 'analogous sorts of social exchanges', which help the robot 'learn the meaning [its] acts have for others' (Breazeal, 1998b, p. 31).

The underlying (apparently Vygotskian) hypothesis about cognitive development is that the combination of 'natural' tendencies, motor acts, and human scaffolding generates intentionality. Thus the combination of endogenous 'smiling' and mother-infant imitation (or mimicry, in the case of the infant) produces voluntary social *smiling*. (On the role of cross-cultural as well as maturational factors in the development of social smiling, see Wörmann et al., 2012; Anisfeld, 1982.) This hypothesis suggests Wittgenstein's own approach to the mind. Nevertheless, employing his criteria for expressive behaviour, it seems that the human

infant does *not* smile. Wittgenstein mocked the idea that a baby might have a normal play of expressions: 'Imagine a newborn child who, of course, couldn't speak, but who had the play of features and gestures of an adult!' (LW I, §945). He also said that the 'newborn child' is not capable of being 'malicious, friendly, or thankful. Thankfulness is only possible if there is already a complicated pattern of behaviour' (LW I, §942). The infant does not have any of the behaviours that are associated with friendliness, or kindness, or malice, and that turn a 'smile' into a *smile*. To describe the infant as *smiling* is to anthropomorphize the child, it seems. This result is in tension with psychological studies suggesting that neonates *do* exhibit social smiles (see Cecchini et al., 2011).[11] It also puts Wittgenstein at odds with the standard view that social smiling emerges by approximately two months of age (for full-term infants),[12] since at this age the infant's behaviour is not yet embedded in the 'bustle of life'.

On Wittgenstein's account, it appears, neither the robot nor the baby smiles. Yet infant 'smiling' behaviour is very different from Kismet's behaviour and is closer to that of a paradigm smiling adult. The infant's 'expressions' have greater variability and irregularity than Kismet's displays; according to Wittgenstein, even a dog is 'more like a human being than a being endowed with a human form, but which behaved "mechanically"' (RPP II, §623). Wittgenstein also emphasized that human behaviour is embedded in multiple, complicated contexts, and the infant's 'smile' is embedded in many more contexts than is Kismet's 'smile'.

To recognize these differences between the child and the 'child-machine', Wittgenstein's remarks about neonates quoted above must be tempered. In fact he allowed that an infant could have some expressive behaviour, namely imperative pointing. He said that if a child 'stares at the thing with eyes wide open, reaches out towards it with his hand, and perhaps lets out a cry', this 'can perfectly well be called an *expression* of a desire' (VW, p. 419). (The child's 'desire' is not a 'conscious' desire; without language to provide the *object* of desire, there is only 'perhaps an agitation, an obscure urge, a feeling of tightness in the chest' (VW, p. 421).) This suggests that an infant's 'smile' *could* be called an expression of pleasure (or friendliness). Wittgenstein also said that 'there is such a thing as "primitive thinking" which is to be described via primitive *behaviour*', and that our language-games are 'an extension of primitive behaviour' (RPP II, §205; Z, §545). Perhaps the child could be said to have 'primitive' emotions which are expressed via primitive expressive behaviour. As usual, we can follow Wittgenstein's injunction to say

'what you choose, so long as it does not prevent you from seeing the facts' (PI, §79).[13]

8.6 The smiling pianola

From Kismet's 'smile' we can deduce information about the robot's internal states. According to Breazeal, 'the robot's outwardly observable behavior must serve as an accurate window to its underlying computational processes' – the states of its 'emotion' and 'drive' systems (2002, p. 10). According to Wittgenstein, in contrast, emotions are *not* inner states. Fright, for example, is not 'something which goes along with the experience of expressing fright'; expressing fright is instead *part* of the fright (LPE, p. 202). One corollary of this is that we 'do not see facial contortions and *make the inference* that he is feeling joy, grief, boredom' (RPP II, §570). We simply *see* emotion in the face:

> Look into someone else's face and see the consciousness in it, and also a particular *shade* of consciousness. You see on it, in it, joy, indifference, interest, excitement, dullness etc. (RPP I, §927)

For Wittgenstein, a facial expression is a (continuous) *aspect* of a face; to see the aspect, we simply '[e]nter into the expression' and let 'the face impress itself' on us (RPP I, §1033; BBB, p. 165). Seeing an aspect 'presupposes concepts which do not belong to the description of the [object] itself', and depends on 'my knowledge, on my general acquaintance with human behaviour' (RPP I, §§1030, 1073). Thus, seeing a *smile* requires concepts such as friendliness rather than merely anatomical concepts. Wittgenstein's criterion for whether one *sees* an expression is how one responds to the face. For example, he said 'Whoever senses [sadness in a face] often imitates the face with his own' (LW I, §746) – this is affective mirroring, in current terminology. The significance for social cognition of automatically-occurring facial movements in response to facial expressions is a much-researched area in recent years (see e.g. Schilbach et al., 2008, on the neural correlates of spontaneous facial reactions to perceived facial expressions). It is striking that Wittgenstein emphasized the importance, in facial-expression processing, of imitating expressions.

Wittgenstein also anticipated the modern discussion of deficits in facial-expression processing – *mind-blindness*, in current terminology. He remarked, 'One might say of someone that he was blind to the *expression* of a face' (LW I, §763; and also, 'I can very well imagine someone

who, while he sees a face extremely accurately, and can, e.g. make an accurate portrait of it, yet doesn't recognize its smiling expression as a smile. I should find it absurd to say that his sight was defective' (RPP I, §1103). The mind-blind person reacts differently to emotional expressions, according to Wittgenstein: someone who 'sees a smile and does not know it for a smile, does not understand it as such ... mimics it differently' (PI II, p. 198). This is consistent with current research on rapid facial responses to emotional expressions that finds differences between the facial reactions of individuals with and without autism spectrum disorders (ASD) (see Beall et al., 2008).

On Wittgenstein's approach to mind-reading, the mind-blind individual does not lack a *theory* of mind, or the ability to *simulate* another mind, but instead suffers from *aspect*-blindness. Aspect-blindness, Wittgenstein said, is like 'the lack of a "musical ear"' (PI II, p. 214). His approach might generate alternative ways of understanding and identifying mind-blindness, and alternative therapies for impaired performance in (affective) facial-expression recognition. Again Wittgenstein's anticipation of current psychological theorizing is striking. Gangopadhyay and Schilbach (2012), for example, propose an embodiment-based alternative to theory-based and simulation-based analyses of mind-reading, and claim that we have 'immediate perceptual access' to other minds – this is analogous to Wittgenstein's claim that we can look 'into someone else's face and see the consciousness in it'.

According to Wittgenstein, the mind-blind individual can identify facial anatomy (LW I, §780) – that is to say, non-affect facial processing may be unimpaired. Moreover, the mind-blind individual may be able to infer the affect expressed by a face from the geometry of the face, or from other people's responses to the face:

> Think of our reactions towards a good photograph, towards the facial expression in the photograph. There might be people who at most saw a kind of diagram in a photograph, as we consider a map; from it we can gather various things about the landscape.... (RPP I, §170)

Such people might lack the concept of (say) a *hesitant* facial-expression, but 'have a concept which was always applicable where "hesitant" ... is' (LW I, §741). The mind-blind person's concept would not be equivalent to the original concept (Wittgenstein said, for example, that a *tender* facial expression cannot be described 'in terms of the distribution of matter in space' (LW I, §954)), and Wittgenstein also suggested that there may be facial expressions for which the mind-blind individual would

be unable to construct a co-extensive concept.[14] The idea that a person with ASD might have a concept co-extensive with a typically-developing human's concept is an intriguing hypothesis about the facial-expression processing strategies of individuals with and without ASD.[15] Again, it is a conceptual claim that forms the basis of an empirically testable hypothesis.

According to Breazeal, robots such as Kismet must be able to 'perceive and interpret the human's emotive expressions', in order to produce appropriate responses (Breazeal, 1999). Kismet has a face detector and is biased (given the appropriate states of its 'drive' system) to attend to human faces. Other robots, such as Hara and Kobayashi's Mark II robot, are reported to be capable of 'real-time, automatic recognition of human facial expressions' (Hara, 2004; Hara and Kobayashi, 1997). Their robot 'will greet smiles with smiles, frowns with frowns', Hara says (in Menzel and D'Aluisio, 2000, p. 73). The robot Leonardo is designed to learn to reproduce perceived facial expressions, and on this basis is said to imitate (rather than merely mimic) facial expressions. This is to help to bootstrap 'early forms of emotional understanding for the robot' – Leonardo, it is claimed, learns the 'affective meaning' of facial expressions (Breazeal et al., 2005).

However, Kismet, along with other robots designed to recognize facial expressions, lacks the 'general acquaintance with human behaviour' that (according to Wittgenstein) is required for aspect-perception. It follows that the facial-expression processing strategies of the robot and the (typically-developing) human being must differ – Kismet is mind-blind, it seems. This raises the question whether the facial-expression processing strategies implemented in social robots are in fact more like those of individuals with ASD. Such strategies include, it is hypothesized, perceiving faces as a combination of (context-independent) discrete features, in a systematic, sequential manner, versus perceiving faces as a (context-dependent) 'gestalt' whole (see e.g. Evers et al., 2011).

Given the fundamental differences between a 'smiling' robot such as Kismet and a paradigm smiling human, it is misleading to describe such machines as 'child-machines', or as 'socially intelligent'. But then how are we to characterize social robots? In this regard, Wittgenstein's notion of a *reading-machine* is helpful (see Proudfoot, 2002, 2009). A typical reading-machine is a pianola (or 'playing-machine'), which, Wittgenstein said, translates marks into sounds by 'reading' the pattern of perforations in the pianola roll (PI, §157; BBB, p. 118). A reading-machine 'reads' only in the sense of 'translating script into

sounds, also of writing according to dictation or of copying in writing a page of print, and suchlike; reading in this sense does not involve any such thing as understanding what you read' (BBB, p. 119).[16] The differences between Kismet and a paradigm smiling human suggest that Kismet is (what I shall call) a *smiling-machine*. (Of course, Kismet also produces 'frowns' and other displays – smiling is simply a representative example.) A smiling-machine smiles only in the sense of producing (anatomical) 'smiles' in response to human-appropriate stimuli. Smiling in this sense does not involve *expressing emotion* or *communicating*.

8.7 Conclusion

Computer scientist Woody Bledsoe famously described a dream 'filled with the wild excitement of seeing a machine act like a human being', and he spoke of the 'yearning' that AI researchers have 'to make machines act in some fundamental ways like people' (1986, p. 57). For many modern researchers, building a robot that 'smiles' and 'frowns' and has a 'happy expression' or 'sorrowful' expression is an essential step to realizing this dream.

Wittgenstein, however, demonstrated the unrecognized philosophical complexities in this plan.[17] His claim that a 'smiling mouth *smiles* only in a human face' elliptically expresses his claim that smiling not only requires the 'normal play' of human facial expressions but also must be embedded in complicated forms of human life. Without these 'surroundings', an anatomical 'smile' is simply not a *smile*. Wittgenstein's account of facial-expression is a storehouse of fascinating ideas for both psychology and AI. It makes clear just how difficult the tasks are for strong face AI, and more generally for cognitive and developmental robotics.

Notes

1. The researchers building these face robots are concerned more with the mechanics of human facial-expression, and modelling *anatomical* smiles, rather than giving the robot a 'motivational' system based upon theories of emotions in humans and other animals (see Hara and Endo, 2000).
2. On 'believable creatures' see Bates (1994).
3. Cynthia Breazeal, quoted in Menzel and D'Aluisio (2000, p. 69).
4. Typically, AI researchers deny that their 'emotional' robots actually have emotions. Parisi and Petrosino are an exception; they claim that their simulated robots '*have* emotions because emotions can be shown to play a well-identified function in what they do' (2010, p. 453).

5. Breazeal distinguishes Kismet's 'emotive expressions', which 'reflec[t] the state of the robot's emotion system' from its 'expressive facial displays', which 'conve[y] social cues during social interactions with people' (2002, p. 158). I shall use the words 'expression' and 'display' interchangeably.
6. LSD, p. 367.
7. Wittgenstein said, for example, 'How would a human body have to act so that one would not be inclined to speak of inner and outer human states? Again and again, I think: "like a machine".' (LW II, p. 66).
8. The assumption that this is solely an *engineering* matter is harmless here, given the additional challenge for AI of the embedding problem.
9. Nor are psychological accompaniments *necessary*; non-genuine ('Duchenne') smiles have the same meaning as genuine smiles, even though *ex hypothesi* the insincere person lacks the supposedly appropriate feeling. In any case, Wittgenstein argued that pleasure is not a sensation, and that emotions are not sensations (RPP I, §800; RPP II, §148).
10. http://www.ai.mit.edu/projects/kismet-new/kismet.html.
11. For the related claim that crying in newborns depends, not only on the infant's psychological state, but on the communicative context, see Cecchini, Lai, and Langher, 2007.
12. See Anisfeld (1982).
13. See G. E. M. Anscombe's article 'Pretending' for this approach to the anthropomorphizing of non-linguistic animals (1958, p. 291).
14. Using an analogy with musical expression, Wittgenstein said that a *tender* expression in music 'can't even be explained by reference to a paradigm, since there are countless ways in which the same piece may be played with genuine expression' (LW I, §954).
15. On the computational modelling of gaze, comparing autistic individuals and controls, see Shic et al. (2006).
16. A reading-machine can be 'living' (PI, §157) – for example, Wittgenstein said that a living calculating-machine can produce proofs of complicated mathematical theorems but cannot figure out 'what change you should get from a shilling for a twopenny bar of chocolate' (LFM, p. 36). Might we then regard the mind-blind *human being* (or even the human infant) as a *living smiling-machine* – a human who produces only (anatomical) 'smiles' and does not express emotion by this behaviour?
17. This chapter was written while I was a Visiting Research Fellow in the Department of Psychology at Georgetown University. I am extremely grateful to the Department, in particular to James Lamiell and Rom Harré, and to Rachel Barr for valuable discussion about imitation in infancy.

References

R. Adolphs (2005) 'Could a Robot have Emotions? Theoretical Perspectives from Social Cognitive Neuroscience' in M. Arbib and J.-M. Fellous (eds) *Who Needs Emotions: The Brain Meets the Robot* (Oxford University Press), 9–28.

E. Anisfeld (1982) 'The Onset of Social Smiling in Preterm and Full-term Infants from Two Ethnic Backgrounds', *Infant Behavior and Development*, 5, 387–95.

G. E. M. Anscombe (1958) 'Pretending', *Proceedings of the Aristotelian Society*, Supplementary Volume 32, 279–94.

L. Aryananda (2006) 'Attending to Learn and Learning to Attend for a Social Robot', *Humanoids '06–2006 IEEE-RAS International Conference on Humanoid Robots*, December 4–6 2006, Genova, Italy, 618–23.

M. A. Arbib and J.-M. Fellous (2004) 'Emotions: From Brain to Robot', *Trends in Cognitive Sciences*, 8(12), 554–61.

J. Bates (1994) 'The Role of Emotion in Believable Agents', *Communications of the ACM*, 37(7), 122–5.

P. M. Beall, E. J. Moody, D. N. McIntosh, S. L. Hepburn, and C. L. Reed (2008) 'Rapid Facial Reactions to Emotional Facial Expressions in Typically Developing Children and Children with Autism Spectrum Disorder', *Journal of Experimental Child Psychology*, 101, 206–23.

K. Berns and J. Hirth (2006) 'Control of Facial Expressions of the Humanoid Robot Head ROMAN', *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 9–15 October 2006, Beijing, China, pp. 3119–24.

W. Bledsoe (1986) 'I Had a Dream: AAAI Presidential Address', 19 August 1985, *AI Magazine*, 7(1), 57–61.

C. Breazeal (1998a) 'Regulating Human-Robot Interaction Using "Emotions", "Drives" and Facial Expressions', *Proceedings of 1998 Autonomous Agents workshop, Agents in Interaction – Acquiring Competence Through Imitation*, Minneapolis, MO, 14–21.

—— (1998b) 'Early Experiments using Motivations to Regulate Human-Robot Interaction', AAAI Technical Report FS-98–03 (Menlo Park, CA: AAAI Press), 31–6.

—— (1999) 'Robot in Society: Friend or Appliance?', *Agents99 Workshop on Emotion-Based Agent Architectures* (Seattle, WA), 18–26.

—— (2000) *Sociable Machines: Expressive Social Exchange between Humans and Robots*, Dissertation submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Doctor Of Science at the Massachusetts Institute of Technology, May 2000.

—— (2001) 'Affective Interaction between Humans and Robots' in J. Kelemen and P. Sosik (eds) *ECAL 2001, LNAI 2159* (Berlin: Springer-Verlag), 582–91.

—— (2002) *Designing Social Robots* (Cambridge, Mass.: MIT Press).

—— (2003a) 'Toward Sociable Robots', *Robotics and Autonomous Systems,* 42, 167–75.

—— (2003b) 'Emotion and Social Humanoid Robots', *International Journal of Human-Computer Studies*, 59, 109–15.

—— (2003c) 'Emotive Qualities in Lip-Synchronized Robot Speech', *Advanced Robotics*, 17(2), 97–113.

—— (2006) 'Human-Robot Partnership', *IEEE Intelligent Systems*, 21(4), 79–81.

—— (2009) 'Role of Expressive Behaviour for Robots that Learn from People', *Philosophical Transactions of The Royal Society, Part B,* 364, 3527–38.

C. Breazeal and L. Aryananda (2002) 'Recognizing Affective Intent in Robot Directed Speech', *Autonomous Robots*, 12(1), 83–104.

C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg (2005) 'Learning from and About Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots', *Artificial Life*, 11(1–2), 31–62.

C. Breazeal and P. Fitzpatrick (2000) 'That Certain Look: Social Amplification of Animate Vision' in *Proceedings of the AAAI Fall Symposium, Socially Intelligent Agents – The Human in the Loop*.

C. Breazeal, J. Gray, and M. Berlin (2009) 'An Embodied Cognition Approach to Mindreading Skills for Socially Intelligent Robots', *The International Journal of Robotics Research*, 28, 656–80.

C. Breazeal and B. Scassellati (2000) 'Infant-like Social Interactions between a Robot and a Human Caretaker', *Adaptive Behavior*, 8(1), 49–74.

—— (2001) 'Challenges in Building Robots That Imitate People' in K. Dautenhahn and C. Nehaniv (eds) *Imitation in Animals and Artifacts* (Cambridge, Mass.: MIT Press).

C. Breazeal and J. Velásquez (1998) 'Toward Teaching a Robot "Infant" Using Emotive Communication Acts', *Proceedings of 1998 Symposium of Adaptive Behavior, Workshop on Socially Situated Intelligence*.

R. A. Brooks (1999) *Cambrian Intelligence: The Early History of the New AI* (Cambridge, Mass.: MIT Press).

R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. M. Williamson (1999) 'The Cog Project: Building a Humanoid Robot' in C. L. Nehaniv (ed.) *Computation for Metaphor, Analogy, and Agents: Lecture Notes in Artificial Intelligence 1562* (Berlin: Springer), 52–87.

M. Cecchini, E. Baroni, C. Di Vito, and C. Lai (2011) 'Smiling in Newborns during Communicative Wake and Active Sleep', *Infant Behavior and Development*, 34, 417–23.

M. Cecchini, C. Lai, and V. Langher (2007) 'Communication and Crying in Newborns', *Infant Behavior and Development*, 30, 655–65.

K. Dautenhahn (2007) 'Socially Intelligent Robots: Dimensions of Human-Robot Interaction', *Philosophical Transactions of The Royal Society*, *Part B*, 362, 679–704.

K. Evers, I. Noens, J. Steyaert, and J. Wagemans (2011) 'Combining Strengths and Weaknesses in Visual Perception of Children with an Autism Spectrum Disorder: Perceptual Matching of Facial Expressions', *Research in Autism Spectrum Disorders*, 5, 1327–42.

J. Fasola, and M. J. Matarić (2012) 'Using Socially Assistive Human-Robot Interaction to Motivate Physical Exercise for Older Adults', *Proceedings of the IEEE*, 100(8), 2512–26.

D. Feil-Seifer, and M. J. Matarić (2011) 'Automated Detection and Classification of Positive vs. Negative Robot Interactions with Children with Autism Using Distance-Based Features' in *Proceedings of the International Conference on Human-Robot Interaction* (HRI'11), 6–9 March 2011, Lausanne, Switzerland, 323–30.

T. Fong, I. Nourbakhsh, and K. Dautenhahn (2003) 'A Survey of Socially Interactive Robots', *Robotics and Autonomous Systems*, 42, 143–66.

N. Gangopadhyay and L. Schilbach (2012) 'Seeing Minds: A Neurophilosophical Investigation of the Role of Perception-Action Coupling in Social Perception', *Social Neuroscience*, 7(4), 410–23.

N. Giullian, D. Ricks, A. Atherton, M. Colton, M. Goodrich, and B. Brinton (2010) 'Detailed Requirements for Robots in Autism Therapy' in *2010 IEEE International Conference on Systems, Man, and Cybernetics*, 10–13 October 2010, Istanbul, Turkey.

K. Gold, and B. Scassellati (2007) 'A Bayesian Robot that Distinguishes "Self" from "Other"', *Proceedings of the 29th Annual Meeting of the Cognitive Science Society (CogSci2007)*.

F. Hara (2004) 'Artificial Emotion of Face Robot through Learning in Communicative Interactions with Human', *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, 20–22 September, Kurashiki, Okayama Japan.

F. Hara, and K. Endo (2000) 'Dynamic Control of Lip-Configuration of a Mouth Robot for Japanese Vowels', *Robotics and Autonomous Systems*, 31, 161–9.

F. Hara, and H. Kobayashi (1997) 'State-of-the Art in Component Technology for an Animated Face Robot – Its Component Technology Development for Interactive Communication with Humans', *Advanced Robotics*, 11(6), 585–604.

J. Haugeland (1985) *Artificial Intelligence: The Very Idea* (Cambridge, Mass.: Bradford Books).

E. S. Kim, and B. Scassellati (2007) 'Learning to Refine Behavior Using Prosodic Feedback', *Proceedings of the 6th IEEE International Conference on Development and Learning* (ICDL 2007), July 2007, London, England.

H. Kozima, C. Nakagawa, and H. Yano (2005) 'Using Robots for the Study of Human Social Development', *AAAI Spring Symposium on Developmental Robotics (DevRob-2005)*, 22 March 2005, Palo Alto, Calif., pp. 111–4.

M. Kunz, K. Prkachin, and S. Lautenbacher (2009) 'The Smile of Pain', *PAIN*, 145, 273–5.

S. Lee, I. Y. Lau, and Y. Hong (2011) 'Effects of Appearance and Functions on Likability and Perceived Occupational Suitability of Robots', *Journal of Cognitive Engineering and Decision Making*, 5(2), 232–50.

C. -Y. Lin, L. -C. Cheng, C. -K. Tseng, H. -Y. Gu, K. -L. Chung, C. -S. Fahn, K. -J. Lu, and C. -C. Chang (2011) 'A Face Robot for Autonomous Simplified Musical Notation Reading and Singing', *Robotics and Autonomous Systems*, 59, 943–53.

P. Menzel, and F. D'Aluiso (2000) *Robo Sapiens: Evolution of a New Species* (Cambridge, Mass.: MIT Press).

M. L. Minsky (1968) *Semantic Information Processing* (Cambridge, Mass: MIT Press).

E. Nagy (2012) 'From Symmetry to Asymmetry? The Development of Smile', *Cortex*, 48, 1064–7.

D. Parisi, and G. Petrosino (2010) 'Robots That have Emotions', *Adaptive Behavior*, 18, 453–69.

D. Proudfoot (2002) 'Wittgenstein's Anticipation of the Chinese Room' in J. M. Preston and M. A. Bishop (eds) *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (Oxford: Oxford University Press), pp. 167–80.

—— (2004a) 'Robots and Rule-following' in C. Teuscher (ed.) *Alan Turing: Life and Legacy of a Great Thinker* (Berlin: Springer-Verlag), 359–79.

—— (2004b) 'The Implications of an Externalist Theory of Rule-Following Behaviour for Robot Cognition', *Minds and Machines*, 14, 283–308.

—— (2009) 'Meaning and Mind: Wittgenstein's Relevance for the "Does Language Shape Thought?" debate', *New Ideas in Psychology*, 27, 163–83.

—— (2011) 'Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game', *Artificial Intelligence*, 175, 950–7.

B. Scassellati (2000) 'How Developmental Psychology and Robotics Complement Each Other', *NSF/DARPA Workshop on Development and Learning*, June 2000, Michigan State University, Lansing, MI.

—— (2001) 'Investigating Models of Social Development Using a Humanoid Robot' in B. Webb and T. Consi (eds) *Biorobotics* (Cambridge, Mass.: MIT Press),

—— (2002) 'Theory of Mind for a Humanoid Robot', *Autonomous Robots,* 12, 13–24.

—— (2005) 'How Social Robots Will Help Us to Diagnose, Treat, and Understand Autism', *12th International Symposium of Robotics Research (ISRR)*, October 2005, San Francisco, Calif.

B. Scassellati, C. Crick, K. Gold, E. Kim, F. Shic, and G. Sun (2006) 'Social Development', *IEEE Computational Intelligence Magazine*, August 2006, 41–7.

L. Schilbach, S. B. Eickhoff, A. Mojzisch, and K. Vogeley (2008) 'What's in a Smile? Neural Correlates of Facial Embodiment during Social Interaction', *Social Neuroscience*, 3(1), 37–50.

J. R. Searle (1980) 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3, 417–56.

—— (1992) *The Rediscovery of the Mind* (Cambridge, Mass.: MIT Press).

S. C. Shapiro (2006) 'Natural-Language-Competent Robots', *IEEE Intelligent Systems*, 21(4), 76–7.

B. Shic, W. Jones, A. Klin, and B. Scassellati (2006) 'Swimming in the Underlying Stream: Computational Models of Gaze in a Comparative Behavioral Analysis of Autism', *28th Annual Conference of the Cognitive Science Society*, Vancouver, 780–5.

B. Shic, and B. Scassellati (2007) 'Pitfalls in The Modelling of Developmental Systems', *International Journal of Humanoid Robotics*, 4(2), 435–54.

G. Sun and B. Scassellati (2005) 'A Fast and Efficient Model for Learning to Reach', *International Journal of Humanoid Robotics*, 2(4), 391–413.

A. Tapus, M. J. Matarić, and B. Scassellati (2007) 'The Grand Challenges in Socially Assistive Robotics', *IEEE Robotics And Automation Magazine*, 4(1), 35–42.

A. M. Turing (1950) 'Computing Machinery and Intelligence', *Mind*, 59, 433–60.

E. Wade, A. Parnandi, R. Mead, and M. J. Matarić (2012) 'Socially Assistive Robotics for Guiding Motor Task Practice', *Journal of Behavioral Robotics*, 2(4), 218–27.

V. Wörmann, M. Holodynski, J. Kärtner, and H. Keller (2012) 'Smiling in Newborns during Communicative Wake and Active Sleep', *Infant Behavior and Development*, 34, 417–23.

9

A Return to 'the Inner' in Social Theory: Archer's 'Internal Conversation'

Wes Sharrock and Leonidas Tsilipakos

9.1 Introduction

Like psychology, sociology is often infused with the idea of transforming itself into a genuine science by redesigning itself to conform to a generic model of what scientific explanation is supposedly like. In recent sociology and social psychology, a 'realist' understanding of science has provided a prominent platform for promoting the renewal of scientific ambitions in the 'behavioural sciences'. As usual, the promised science does not soon materialize; its development is postponed until the problems involved in setting out how the proper form of scientific explanation actually works are sorted out. These conceptual problems involve not only issues about the nature of explanation, but also those found in philosophical psychology in relation to how conduct is to be brought within the explanatory scheme: how are the relations between 'the inner' and 'the outer', the 'private' and 'the public' to be encompassed within the scheme? These questions enter, no less awkwardly into sociological as they do into psychological deliberation, often in contexts bounded by the perceived opposition between mentalism and behaviourism, on the one hand, and between determinism and autonomy on the other. In sociology, one setting for these difficulties is the perennial controversy over the proportional contributions to behavioural outcomes made by 'structure' (the influence of pre-existing conditions) and 'agency' (individual spontaneity), where there are two (at least) contested inclinations involved in attempts to foreclose the difficulties: one, to doubt that individual decisions genuinely fix outcomes, and, two, to assume a 'behaviourist' conception of the mind as an exhaustively public

phenomenon. Here we consider a 'critical realist' candidate for the solution of the structure and agency problem, which is the identification of an 'inner conversation' through which individuals respond to given circumstances, but do so deliberatively and decisively. If this idea is to do the work intended for it, then it needs to place itself in relation to the received contrasts of inner and outer, of public and private, but it is in these connections that the whole idea runs into insuperable problems.

Since the time of Marx, at the very least, social theory has concerned itself with what is nowadays commonly referred to as 'the problem of structure and agency'. Phrased as a singular problem about 'structure' and 'agency', it informs the professional reasoning of sociologists who frequently distinguish between theories based on whether they emphasize one at the expense – even to the complete exclusion – of the other. Accordingly, theorists attempting to deal with the problem address it in the way one would a rigid dualism, proposing ways to overcome it via some reconciliation of the terms (e.g. Giddens, 1984). Accepting a problem as faithfully identified by its appellation is, nevertheless, not always the best strategy. For 'the problem' of structure and agency is, really, a rather intractable assortment of *a number of problems*. Some possess a long philosophical pedigree, such as the problem of free will and determinism, and the 'theory' of causality, while others have shorter histories, such as the 'theory' of scientific explanation. These problems are fused together with uncontroversial statements akin to Marx's ('Men make their own history, but … they do not make it under circumstances chosen by themselves' 1977[1852], p. 300), together with an *a priori* commitment to the necessary existence of the social sciences as uniquely capable of (truly) explaining 'social action'. Theorists who attempt to tackle 'the problem' can only address some partial configuration of these issues. 'The problem', thus, changes with every attempt to solve it, although, as noted, it continues to be understood as 'the same' problem of structure and agency.

This chapter examines but *one* form of the alleged problem as it pertains to Margaret Archer's work (2003, 2007), which constitutes one of the latest attempts to settle 'structure and agency', notable for the purposes of the present volume for its appeal to the psychological workings of an 'internal conversation'. Archer is a critical realist, which means that she is, broadly speaking, following Roy Bhaskar's realist theory of science (1975) via its application to the social sciences (1998). Succinctly put, Bhaskar's project can be seen as starting from the idea that science seeks to identify not relations between events, but the tendencies of powerful particulars as bestowed on them by their constitutive mechanisms (Harré and

Madden, 1975). These causal powers are thought to be positioned in open systems and thus shape events non-deterministically. Finally, it is argued that since some powers may exist unexercised or when exercised may have unperceived effects, reality must be seen to be stratified, and unobservable mechanisms lie at its deepest level.[1]

Transferring these ideas to the social domain, the critical realist paradigm assumes that the possibility of a social science depends on the adoption of the proper form of (scientific) explanation, namely explanation in terms of underlying mechanism(s). Following this programmatic commitment, it is thought that structure and agency need to be related in mechanistic terms. A complication arises, however, because the *causal powers* of structure and agency are held to be different *in kind* and, hence, are thought to require some kind of mediation[2] (2003, p. 15), if they are, indeed, to be connected in the form of (and as parts of) action-producing mechanisms. Archer's work attempts to address precisely this difficulty. In what follows, we first exposit her proposed solution in some detail and subsequently subject it to scrutiny.

9.1.1 Exposition

For Archer, part of the answer to these social theoretical troubles has lain – though largely unnoticed – before our eyes in the undertheorized area she marks with the term 'reflexivity'. Reflexivity is meant to refer to a familiar ability (and exercise thereof):

> *'reflexivity' is the regular exercise of the mental ability, shared by all normal people, to consider themselves in relation to their (social) contexts and vice versa.* (Archer, 2007, p. 4, original italics)

Because it is ubiquitous, reflexivity comes in a number of forms,[3] most of which manifest themselves through our equally ubiquitous 'internal conversations':

> At its most basic, reflexivity rests on the fact that all normal people talk to themselves within their own heads, usually silently and usually from an early age. In the present book this mental activity is called 'internal conversation' but, in the relatively sparse literature available, it is also known inter alia as 'self-talk', 'intra-communication', 'musement', 'inner dialogue' and 'rumination'. (Archer, 2007, p. 4)

Archer thinks that 'internal conversation' holds the key to the required mediation between personal and social powers, that it is the

unappreciated inner process which, by connecting personal concerns to one's circumstances,[4] completes the works of the mechanism and effects mediation of structure and agency through provision of a space in which individuals can make decisions, involving the explicit consideration of structural constraints. This process can occasionally misfire or fail to lead to decisive action for the people Archer calls 'fractured reflexives' (2003, p. 164), but in principle it bears the deliberative burden and fulfils the necessary if it is to bear the deliberative burden and, if it is to be part of a mechanism, fulfills the necessary criterion of being causally adequate of being causally efficacious:

> Inner conversations … can lead to the re-setting of watches with the (putative) correct date; the continued writing of this chapter and to our determined commitment to life-projects'. (Archer, 2003, p. 105)

Although 'internal conversation' can appear in scare quotes in Archer's work, internal conversation is not 'internal' conversation, and it is not internal 'conversation' either. Firstly, 'internal conversation' is held to be genuinely interior (2003, p. 16). Secondly, the impression that conversations we have with other people and internal conversation are two species of the same genus is upheld. There might, indeed, be differences, such as the following:

> We cannot offend ourselves (though we may later deem our thoughts to have been offensive), we can interrupt ourselves as often as we please, start dialogues where we wish and abandon them when we will, we never apologise for dwelling interminably on some preoccupation or breaking off in mid-sentence, and we can curse our heads off or bawl our hearts out – all because questions of self-presentation, public accountability or consideration for others do not arise. (Archer, 2007, p. 74)

However, the above distinctive features are not thought to militate against the applicability of the label conversation to our inner workings. Besides, 'internal conversation' involves interlocutors – as any conversation should (2003, pp. 70–1) – though not in the sense of different people, but, rather, of analytically distinguished aspects of the same person addressing one another:

> I only talk to myself and the internal conversation is not between three reified people inside me. (Archer, 2003, p. 75)

> I do nothing other than speak to myself. All the previous considerations pertain to how I do it and not to who is conversing with whom. 'I says to myself says I' cannot be bettered, although it has needed extending to include 'I responds to myself responds I'. (Archer, 2003, p. 107)

The last sentence indicates why 'internal conversation' is not a monologue, the difference being marked by the presence of an internal response (2003, p. 105).[5] Archer phrases this as the thesis that '[W]e can and do speak to ourselves' (2003, p. 44) – with the emphasis (also) on *ourselves* (2003, p. 90). Moreover, according to Archer, 'internal conversation' takes the form of a question-and-answer exchange where there is asking and responding. The question-and-answer format is why 'internal conversation' is held to be capable of *replacing* the erroneous notion of introspection in an account of how we arrive at self-knowledge.[6] In this respect, Archer relies on Gerald Myers (1986), who, let it be noted, considers the following as *a version of* introspection:

> [The importance of] self-dialogue and its role in the acquisition of self-knowledge, I believe, can hardly be exaggerated. That it plays such a role is a consequence of a human characteristic that deserves to be judged remarkable. This is the susceptibility of our mind/body complexes to respond to the questions that we put to ourselves, to create special states of consciousness through merely raising a question. It is only slightly less remarkable that these states provoked into existence by our questions about ourselves quite often supply the materials for accurate answers to those same questions. (Quoted in Archer, 2007, p. 3)

Archer follows this quotation with the thought that:

> Precisely because our reflexive deliberations about social matters take this 'question and answer' format, it is appropriate to consider reflexivity as being exercised through internal conversation. (Archer, 2007, p. 3)

Thus, 'internal conversation' is the form *par excellence* suited to reflexive conduct (and fitted its intermediary role, since it is exempt from 'external' constraint), and as such, what for Archer can and does, indeed, bear the weight of mediating between ourselves and our environment:

> The key feature of reflexive inner dialogue is silently to pose questions to ourselves and to answer them, to speculate about ourselves,

> any aspect of our environment and, above all, about the relationship between them. (Archer, 2007, p. 63)

This concludes our brief exposition of Archer's account.

9.2 The conscription of ordinary facts and concepts

In sketching Archer's conception of 'internal conversation' and her hopes about what it can do for social theory, we have mostly refrained from criticism. It seems to us that she is dealing at the most basic level with the fact that people do a number of things[7] in their everyday lives, a fact which heretofore we have most likely considered as a platitude but which now acquires significance by being attached to a theoretical problem. That we sometimes consider or deliberate on ourselves in relation to our circumstances is used, under this description, as a rigid form of words capturing the reflexive structure of our doings: a logical relationship is stipulated between doing all these things and being reflexive, where 'ourselves' and our 'circumstances' or 'environment' are partitioned between 'agency' and 'structure'. Thus, things that people can uncontroversially be said to do are conscripted to Archer's project which attempts to present them – under the name 'internal conversation' and represented as realizations of the convoluted process of 'reflexivity' – as if they involved any new understanding of how conduct comes about.

The fact that Archer utilizes what are anybody's ordinary doings, might seem to make pinning down her position rather tricky, for she can fall back on the uncontroversial at will. However, we will show that these doings are not properly portrayed by Archer and that, further, her construal of ordinary concepts which apply to such doings are considered on the basis of those surface grammatical features that fit her theoretical picture.

It is worth stressing at this point that what role exactly these concepts play is not open to view. To see this, consider first how the concept of 'internal conversation' is explicated when put to the research subjects she interviewed:

> the following question was used during the pilot investigation: 'Some of us are aware that we are having a conversation with ourselves, silently in our heads. We might just call this "thinking things over". Is this the case for you?' Since all subjects answered 'yes' and no subject in the pilot study expressed difficulties with the substance

> of the question or its formulation, it was retained unchanged in the main investigation. (Archer, 2007, p. 91)

True to the conscription pattern we have already hinted at, the concept in question is presented together with an ordinary formulation of what anybody does. Thus, far from subjects' responses constituting evidence for their engaging in 'internal conversation' possessing the full-fledged features Archer ascribes to it, what is being agreed to by one who assents to the above formulation is only that they sometimes think things over, which can be described in *some* sense as at least partially involving 'internal conversation'. It is important, however, to get right the logical order of what explicates what, for it is what we call thinking things over which gives sense and familiarity to the notion of 'internal conversation' and not vice versa.

The question of the role of vernacular concepts also crops up in the following passage. Archer thinks she has discovered an internal phenomenon which, besides being something professional sociologists have neglected to study, also resists penetration from (some of) our available descriptions.

> Some of the subjects interviewed and also certain social psychologists, respond in a derogatory manner to the idea of 'talking to oneself'. Indeed, this is probably the worst vernacular formulation through which to ascertain anything about their internal conversations from the population at large. (Archer, 2007, p. 4)

Indeed, Archer needs to separate 'internal conversation' from the description in question if she is to defeat someone's disavowal and insist that still they engage in 'internal conversation'. But this move pretends as if Archer relied on anything other than the forms of words that her research subjects use to describe what it is they were doing and, sure enough, it is these vernacular concepts of doings that she uses herself. For Archer, and this is the main contention of her work, such doings mediating between 'agency' and 'structure' receive a somewhat determinate form: they are internal conversations comprising question-and-answer sequences. To repeat, however, the fact that mostly anybody can do such things as reflect or imagine. is not evidence *for* the claim that we do them *in the form* of internal conversations, although some of these concepts seem to suggest 'inner doings' more than others do (e.g. rehearsing silently for a part in theatre and imagining a purple rhinoceros rather than, say, imagining a world without poverty).

In the remainder of this chapter, we will sceptically examine the main features of 'internal conversation' with regard to the following two kinds of considerations:

The first pertains to the character of the 'internal conversation' and to whether the phenomena Archer describes persuasively qualify as the constituents of an ubiquitous 'internal conversation' – perhaps Archer's inner conversation is only an attempt to selectively focus on and exaggerate the significance of some features of what is ordinarily called 'thinking', especially under the name of 'reflexivity' of those exercises which are called 'reflection'. The substantive part of the discussion will attempt to breathe life into some of the key forms of words that Archer employs abstractly throughout her books (e.g. 'talking to oneself', 'we speak to ourselves', 'questioning and answering ourselves'), by embedding them into situations where they are actually used. We intend to examine whether she gets these expressions right and whether they do indeed lend any support to her thesis.

The second consideration pertains to the status of 'internal conversation', whether it can occupy the special position Archer would assign it in the understanding of conduct. We examine her potential rationale for resorting to 'the inner' and, finally, we assess 'internal conversation' as a candidate solution to the 'problem of structure and agency'.

9.3 Is 'internal conversation' ubiquitous?

Without doubting that people can think things without announcing them we can nonetheless view the idea that doing this comprises a psychological process of 'internal conversation' as lacking in plausibility. We recognize, however, that one of the ways that this idea might suggest itself is, apart from being associated with the aforementioned everyday activities anybody engages in, if 'internal conversation' were to be identified with thinking and 'thought processes', themselves thought of as omnipresent.

Now it is curious that Archer speaks of 'internal conversation' as if it were not just dialogue but rather a dialectic, a question-and-answer sequence probing ourselves for (self-)knowledge, because doing so undermines the pervasive nature of the process: Simply put, not everybody can be Socrates. By following this line of thinking, however, examples spring to mind of other philosophers who seem to have engaged in something akin to *internal* Socratic dialectics: One may recall Augustine's soliloquizing and, closer to home, Wittgenstein's notebooks.[8] In order to appeal to those examples, one would have to maintain that the writing

consists in reporting an occurring inner dialogue rather than being a form of thinking on its own. However, as far as Augustine is concerned, for instance, there is at the time of his writing a philosophical genre he can be seen to respond to (cf. Stock, 2010). The important thing to take into account is that these forms of writing, much like, for example, the technique of interior monologue as used in novels are indeed specialized literary forms that do not serve in any straightforward sense as ways of representing and reporting inner happenings of daily life.

Whence then the question-and-answer form? One partial answer, for now, is that Archer's theoretical picture requires inner reflection to take place if people are to actively shape their own lives (2003, p. 116) and, if it is to be a conversation, it must consist in related linguistic forms that can be assigned to differentiated interlocutors. Is Archer right, however, to portray such reflection as a conversation genuinely possessing a question-and-answer form?

9.4 Does 'internal conversation' take the form of a question-and-answer exchange?

Archer's idea that our inner workings take the form of a conversation that comprises questions and answers aiming at self-knowledge is neatly summarized in the following passage. The examples invoked are also worth our attention:

> If we review the tasks undertaken by reflexivity or the functions of reflexive processes, we come up with that long list of mental activities bearing the 'self-' prefix: self-observation, self-monitoring, self-criticism, self-evaluation, self-commitment and so forth. All of these have a common denominator. In each and every such activity, we are asking ourselves questions. In everyday language, their respective exemplars are 'How do I look?', 'Am I getting this right?', 'Can't you be more exact?', 'Why did you slip up there?' or 'Could I really do that? 'And to each of these kinds of questions we give ourselves answers, fallible as they are. In other words, by questioning and answering we are holding an internal conversation with ourselves and inter alia about ourselves. (Archer, 2007, pp. 72–3)

Archer thinks that 'asking ourselves questions' – in the sense of (silently) forming interrogatives – is what is common to these 'mental activities', and she proceeds to offer exemplary questions which we can respond to fallibly. Let us explore this idea by reminding ourselves at some length of

cases where 'asking ourselves' and other cognate expressions that Archer employs are, in fact, used. Differences between doings directed at us and at others surface in the following, together with other relevant features:

Talking to oneself

a) We say that one has the habit of talking to oneself, for instance, when walking down the street. This is something one can equally well do out loud: one who talks to oneself out loud while walking may startle others.
b) Imagine the following situations:
   i) Mary is on the phone. Jim enters signalling to her. Placing her palm on the handset she whispers: 'I'm talking to John'.
   ii) Mary and John are having a conversation. Mary wants to know what John thinks about British politics. Jim interrupts but Mary will have none of it: 'I'm talking to John'.
   iii) Jim is in the kitchen washing the dishes. He notices a new saucer. He utters, 'Mm, I've never seen this one before'. Mary, who is in the living room, thinks she heard something. She shouts, 'What did you say?' to which Jim responds, 'I'm talking to myself.'

   In the above situations, 'I'm talking to John/myself' does not characterize or report any question/answer sequence. In the latter two cases, for example, it serves to clarify the intended recipient of an utterance or lack thereof. Again, 'talking to myself' involves a prior audible utterance.
c) It makes sense to talk to Mary to find out whether she knows. But 'I'll talk to myself to find out whether I know' is absurd. Talking to Mary can be informative, talking to myself cannot.

Asking oneself

a) Mary and Jim are having a fight. One of them says to the other: 'I just don't know, I ask myself whether this relationship is going anywhere'. The other responds, 'Well, ask yourself this: whether we have really spent any time together this week'. An interactionally apposite response, one might say, crafted to echo the first utterance. In this case, the person troubled was not waiting for themselves to respond to their second thoughts, but rather making them known to their partner.
b) A novice at chess asks her coach how to decide which move to play.[9] The coach responds: 'First you ask yourself whether your opponent

is threatening something. Then you ask yourself whether any of your pieces are unprotected or any move you intend to make will remove a defender'. This sets out the procedure to be followed, the steps one goes through. One may ask oneself, but one solves the problem by looking at the board. 'But blindfold-chess players ask themselves and look to themselves!' it may be objected. One can play blindfolded, but doing so is parasitic on playing with the board, in that one tries to visualise the board. It is also important that a third member is using a board or some other means of recording the progress of the game, for players may make mistakes as to the position of the pieces.

When one of us who plays chess does not see what he should play, he resorts to asking himself: 'How can I improve my position?' He keeps asking the *question* as long as he is searching for a move to make. And then the *answer* comes in the form of finding a suitable move. This seems to suggest a question–and- answer sequence. But why do we call finding the move an answer here? Is it because he responds inwardly 'by playing Rook to a3'? Rather, we would suggest, if we can call what precedes a *question*, then what properly follows can also be called an *answer*. We can also speak of an answer because that is what stops one asking. Problem and solution, question-and-answer, become interchangeable in this case. Answering the question means working out the solution to the problem.

The player who has a decisive advantage in a game should remind himself: 'There is no need to rush'. He could also say, 'Do I need to rush?' 'No'. In the chess novice case, the coach could also have said, 'First you look for threats'.

c) At times of trouble, one may confide in a friend: 'I ask myself whether I will ever be able to live my life the way I want to'. It might be open to the other person to respond by 'And what does yourself respond?' but that is heard either as a jocular request to report – *per impossibile* – the response one gave or as an attempt to get the person who obviously longs for this not to stop hoping.

Questioning oneself and one's abilities

a) To question someone typically involves asking someone questions as does, for example, courtroom questioning. Similarly, Socrates' questioning of his interlocutors also involves many questions (though not exclusively). On the other hand, questioning oneself is not asking oneself questions but rather closer to doubting one's abilities.

b) Questioning or doubting someone's ability to do something has a much weaker connection to asking questions and can be manifested in, say, disbelief in the face of reassurances offered and refusal to delegate a task in question to that person. Someone who is sceptical about their own abilities may receive the sympathetic response: 'Stop questioning yourself.'

'I say to myself'

(a) 'I say to myself', like 'I said to myself', is used while telling a story as a narrative device to report one's thoughts. 'When the airport staff informed us that there was a two hour delay, I said to myself: "That is it, I'm never going to make my connecting flight"'. This is not followed by any response.

It can be argued, based on the above, that the concept of a conversation does not apply in the same way both to what we do with other people and to what we can call 'talking to oneself' or 'asking oneself'. Moreover, in none of the cases we have examined is Archer's conception of an 'internal conversation' comprising sequences of questions and answers supported. In other words, even if some of the things we say can be positioned within such sequences as potential questions, the question form is superficial (cf. PI, §21) to what someone who might be using these words could be saying, in that it is not a good indication as to what the speaker is doing. This is crucially the case regarding the questions Archer suggests as exemplary:

> 'How do I look?', 'Am I getting this right?', 'Can't you be more exact?', 'Why did you slip up there?' or 'Could I really do that?' (Archer, 2007, p. 72)
>
> questions we put to ourselves, even if we cannot supply the answers: 'Why does he always rub me up the wrong way?', Why do I often type 'becuase' instead of 'because', or 'why did I believe I wouldn't need a jersey today? (Archer, 2003, p. 255)

Most of these questions, rather than being difficult to answer, do not call for an answer in the first place and are instead ways of, among other things, prompting oneself to be careful, complaining or expressing doubt about oneself. Even if we were to accept, nevertheless, that asking these questions can receive an answer (in some sense and given the latitude of 'answer' as shown by previous examples), it is still puzzling how

questions such as 'How do I look?' and 'Why does he always rub me up the wrong way?' are directed at ourselves in such a way so as to produce self-knowledge (unless 'self-knowledge' is construed to consist in being able to give an answer – a correct one – to any question that one is asked about oneself, in which case being able to give one's name might be included). For example, by asking the first question in front of a mirror, we are not addressing ourselves in the way that someone who does not know is addressing someone who knows. We decide whether we look ok, or great, or terrible by looking at the mirror. This, of course, can sometimes be done at a glance and without any inner exchange. Archer seems to think that we always decide by doing something else, where 'by' is read as referring to a process, which, for Archer, is the 'internal conversation'.

This idea, together with Myers' assertion – quoted in a previous section – that 'self-dialogue' is an improved version of introspection, proposes a role for 'internal conversation' as crucial to self-knowledge. Having argued that the idea that we know things about ourselves through introspection is flawed, Archer intends to reserve the selfsame function for 'internal conversation'. Thus, the 'internal conversation' is taken to be the mechanism through which we produce, rather than discover, knowledge of ourselves, much like in the case of introspection, through some internal process.[10] We will have more to say about the idea that 'internal conversation' contributes to self-knowledge. For now, given that the conceptual territory around 'self-knowledge' is quite complex, it may suffice to remind ourselves of some relevant expressions.

To know oneself

'Know thyself' can be heard as a *riposte* to boasting, over-ambition and arrogance. (Perhaps this was its intended meaning as an inscription in the Temple of Apollo at Delphi.) It can also be heard as an exhortation to devote time to exploring understanding and developing one's aptitudes and inclinations. One can reply to a suggestion for an activity at odds with one's character thus: 'I know myself; I would never be able to like this'.

To have self-knowledge

One who has self-knowledge is not surprised by their feelings or reactions, can project how they would behave in a novel situation, claims an accurate assessment of their competences and preferences, perhaps also of how other people assess them.

Having questioned whether the connection between self-knowledge and the inner conversation is as clear as Archer seems to imagine, we can suggest that it is not any precise, even extensive, exploration of what might comprise self-knowledge and how it might be acquired that drives Archer's advocacy. It is, rather, the prior commitment to a theorized picture dictated by 'structure and agency'. It is this picture which also provides a role for 'the inner'. For the structure of the overarching 'problem of structure and agency', that is to say the identification of agency with 'ourselves' and structure with 'environment' and 'circumstances' favours the further alignment of the terms with 'the inner' and 'the outer' and thus manufactures the opposition between the two. We then get either of two possible theoretical pictures: either agency is located at the level of action which is the product of internal deliberation where structural constraints are taken into account before one acts, or alternatively, what goes on internally is the process of agency, in other words our inner space is the space of freedom, which implies that the moment something becomes 'external' it ceases to be absolutely free. This seems to identify circumstances as external constraints and to render impossible the idea that we can be *free in certain circumstances* by construing the latter as some kind of oppressive container. Of the two pictures on offer, Archer, it appears, has opted for the latter.

9.5 A solution to structure and agency?

In light of what has been argued so far, it is hard to see how the 'internal conversation' could contribute to solving the structure and agency problems. Archer's conception of a solution is shaped by her (critical realist) suppositions about the form of (sociological) explanation, which involves identifying the causal connections between two independently identifiable entities with their intrinsic causal powers. Those same suppositions favour a search for black boxes to open up. If what goes on inside the head is a hidden mechanism, we must somehow uncover its workings; searching for and prioritizing 'the inner' is thus a natural move. This explains, in part, the attraction of Archer's picture. The more pressing need, however, in respect of structure and agency lies in establishing the nature of those causal powers exclusive to the agent, since there are strong tendencies within sociology to obliterate these powers. Thus, the 'internal conversation' can be offered as something that belongs entirely to the agent which, in its potential for retention as a wholly private occurrence, is discrete from the public transactions within which the agent is inserted.

Relative to the (various) issues involved in sociology's structure and agency debates, the critical realist rationale is merely a detour which only begs questions about the extent to which conduct is independent of social structures, the extent to which it is authentically under the agent's control. Does the agent exercise any autonomy, or is conduct rather the outcome of pressures originating within social structures?

The 'internal conversation' is meant to show that individuals have discretion in their reaction to the circumstances that social structures generate, enabling the individual to appraise those circumstances, deliberate on possible responses and execute any decision issued from the deliberation. One can certainly agree that conduct is not an automatic response to circumstances, and that individuals can consider and act on their review of the circumstances that confront them, but this does nothing to advance the argument against any determinist view of the effect of social structures. These latter views can feature the recognition that individuals ostensibly deliberate and decide along with the doubt that these are genuinely effective. It is then simply a presupposition of Archer's argument that the affairs of the 'internal conversation' introduce autonomy and executive capacity into the story for they rather presuppose it – the 'internal conversation's' dialogical exchanges only display powers of agency by virtue of their deriving from the general capacities for spontaneity, deliberation and execution that people more generally have, ones which they can exercise throughout their doings and not just in and through those which take place in privatized reflection. If one is prepared to cede autonomy and executive powers to the agent, there is no need to postulate the 'internal conversation' as the source of these, for the capacities manifested in the 'internal conversation' – to ask and answer questions, to review one's situation, imagine, plan and execute a response to them – are only further manifestations of the agents general abilities to do those things, whether *sotto voce* or otherwise.

Though plainly displaying its disagreement with structural determinist views, Archer's proposal does nothing effective to counter them. However, Archer has other reasons for requiring the 'internal conversation', which arises from the need to counter behaviourist tendencies which she understands to entirely empty out the inner. The behaviourism she responds to supposes that all that is to be known about people is to be known by observation, with the result that self-knowledge is only a species of what any observer can know.[11] Using 'self-knowledge' very broadly, to subsume the array of things that under some circumstances one might be said to know about oneself,

it is clear that this idea is flawed, since there are many things one knows about oneself without observation – one's birth date, where one's foot is, what one thinks of the coalition government. Whilst one may know these things without observation, it does not follow that one knows them by a different method than observation 'no method is involved' or both. Self-knowledge is not exhausted by those items which are available to first-person authority, and there are certainly things about oneself that can be opened to scrutiny – is one as upright a person as one takes oneself to be? What method is going to be used in such a reflective inquiry? Even if one proceeds after the fashion of an 'internal conversation', one is not engaged in acquiring information by interrogating a better-informed source (one's analytically distinguishable dialogical counterpart), nor is one, by means of the question-and-answer exchange, reviewing any range of inner psychological or mental conditions whose discernible configuration determine one's character. Of course, individuals may reflect on their possession or lack of personal characteristics, but the basis for that assessment must, of course, be the same one as it is for others, namely the track record. Whether they genuinely are as they might wish to be involves appraisal of how they have acted, just the same basis on which third persons might confidently (and possibly more accurately) make assessments of their character. The difference between first- and third-person assessments in such cases is not in their respective methods, but in the 'database' available. Others are relatively disadvantaged in such cases not by their lack of access to the 'internal conversation' but by the more limited number and range of instances of relevant conduct that they have access to. One is perforce present at all the instances of one's own behaviour. In this sense, first- and third-person assessments are the same, made by reviewing relevant (characteristically recollected) cases.

Archer's concentration on rebutting a behaviourism that denies the reality or relevance of 'inward' occurrences orients her arguments in the wrong direction and ensures a continuing attachment to the other extreme, the idea that there is something distinctive to the inner, as though it were discontinuous with the outer, as though there were different kinds of operations taking place in silent reflection to those taking place in conduct. Breaking from this conviction does not involve denying that people can keep their thoughts to themselves. The fact that they can engage in solitary imaginings and silent decision making does not make the 'internal conversation' *the* site for the reflective and decisive powers of agents; in those respects, it is

only on a par with more overt doings. The latter also involve powers of reflection, deliberation and decision, but not by virtue of the fact that people can think things over in privacy. The idea of the 'internal conversation' makes no additional or especially telling contribution to resistance to ideas of structural determinism. The fact that such an idea is significantly motivated by the prior espousal of a mechanistic explanatory scheme supposedly possessing scientific credentials should demonstrate that, not only sociology and psychology, but the sum of the 'behavioural sciences' ought to be much more cautious about embracing such schemes, especially in respect of expecting substantial enhancement of actual explanatory success from their application to actual cases.

Notes

1. Some of these moves may be questionable, but we will not subject them to scrutiny here (see Hacker, 2007, ch. 4).
2. It is not clear that this proposition actually follows regardless of how one conceives of the distinction in kind. Crudely, the implication hinges on the idea of a mechanism: If structure and agency are somehow parts of a mechanism producing action, there is then a need to explain how these parts of the mechanism are connected. However, if the distinction in question is that between what can be said of social structures and what can be said of individuals, i.e. a conceptual one, then no mechanism is involved, and mediation can only mean a stipulation to change or erase some of the distinctions that we make.
3. Archer distinguishes among 'communicative reflexives', 'autonomous reflexives', 'meta-reflexives' and 'fractured reflexives'. See Archer (2003, 2007).
4. Consider the following excerpts: '… the full mediatory mechanism has been held to depend upon human reflexivity; namely, our power to deliberate internally upon what to do in situations that were not of our making' (Archer, 2003, p. 342). 'Without attending to this mediatory mechanism, which is the internal dialogue, it is impossible to grasp *how* the individual can be an active subject in shaping his or her own life' (Archer, 2003, p. 116).
5. According to Archer, the idea is partly drawn from Peirce and involves speaking, listening and responding as constituting the life of the mind (2003, p. 66).
6. 'Our interior dialogues are not matters into which we *spect intra*, but rather conversations to which we are party, by speaking and by "listening-in"' (Archer, 2003, p. 56). Hacker (this volume, p. 15) reminds us that 'there is such a thing as introspection' but warns that 'it is not a kind of inner sense (or 'apperception' as Leibniz denominated it). It is either a form of self-reflection characteristic of introspective personalities like Proust, or a matter of registering how things are with one (as when one keeps a diary of one's pains for the doctor).'
7. The list of 'mental activities' that was, according to Archer's own account, put before her interviewees is the following: 'mulling over' (a problem, situation

or relationship…), 'planning' (the day, the week or further ahead…), 'Imagining' (as in 'What would happen if…?'), 'deciding' (debating what to do or what's for the best…), 'rehearsing' (practising what to say or do…), 'reliving' (some event, episode or relationship…), 'prioritizing' (working out what matters to you most…), 'imaginary conversations' (with people you know, have known or know about…), 'budgeting' (working out if you can afford to buy or to do something, in terms of money, time or effort…) and 'clarifying' (sorting out what you think about some issue, person or problem…) (2007, p. 91).

8. 'Almost the whole time I am writing conversations with myself. Things I say to myself tête-à-tête'. MS 137; MS 134b, p. 88.
9. The chess player might not qualify as a Rylean 'La Penseuse', but she definitely qualifies for Archer's social actor reflecting on the relation between self and circumstances.
10. 'Self-knowledge is something that we *produce* internally and dialogically; it is not something that we *discover* 'lying inside us' (Archer, 2003, p. 103).
11. In arguing against an extreme, at least, a peculiar, kind of behaviourism, Archer is apt to treat 'inner' and 'private' as though they were equivalent expressions, and as though the everyday notion of 'private' (excluding others) were the same as the philosophical notion of private, one of logical inaccessibility. Perhaps it is the idea that sociology is a purportedly empirical affair that seduces Archer into combining two distinct positions, one pertaining to a rule of investigative method, the other to a rule of grammar. Sociology deeply and extensively conceives itself as an observational discipline (whether it is what it imagines, and even prides itself on, is another matter), and insofar as it does so, then the ascription of beliefs, and so forth. is understood as an evidential affair, which means, of course, that the circumspect investigator will only attribute those attributes for which there is evidence – hence, the ascription of beliefs takes place on the basis of manifestations of them that are public to the observer (we do not either say that this rule is observed in practice). *In that context,* the only thoughts, beliefs, and so forth. that the subjects of inquiry can possess are those that are manifest to the *sociological* observer. Such a position does not deny that persons can have thoughts and beliefs that have not been manifested but only excludes those from consideration, since there are no (evidential) means for determining their identity. The other position, of course, pertains to rules of grammar, to what it is intelligible to say, and that certainly does not exclude recognition of the possibility that persons have thoughts, beliefs, and so forth. which are not made public. It is perfectly intelligible to wonder what someone else is thinking, to anticipate that they have, say, views on, or reactions to, a matter though they have given no expression to them. The use of 'private' here does not indicate something which it logically inaccessible, only something which is contingently and relatively inaccessible. Someone's thoughts, opinions, and so forth. can be kept private, but it is not necessary that they take an 'inner' form, one which they are destined to take, because they cannot be made public in any form, it is only that they have not been manifested. 'Private' does not, either, entail 'inner', since their privacy refers only to their being kept from others, something which can be achieved not only by keeping inner thoughts unannounced but by expressing those thoughts in

spoken soliloquy or by confiding them to one's locked and concealed journals. It is a feature of Archer's own examples of the inner monologue that they are not necessarily exclusively inner – someone who asks themselves, 'Where are my keys?' can 'answer' themselves not by silently soliloquizing to themselves, 'They might be in the kitchen' but by going and looking on the kitchen working top and saying out loud, 'I thought I'd left them there'.

References

M. Archer (2003) *Structure, Agency, and the Internal Conversation* (Cambridge: Cambridge University Press).

—— (2007) *Making Our Way through the World: Human Reflexivity and Social Mobility* (Cambridge: Cambridge University Press).

R. Bhaskar (1975) *A Realist Theory of Science* (Leeds: Leeds Books).

—— (1998) *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences*, 3rd edn (London: Routledge).

A. Giddens (1984) *The Constitution of Society*: *Outline of the Theory of Structuration* (Cambridge: Polity).

P. M. S. Hacker (2007) *Human Nature: The Categorial Framework* (Oxford: Blackwell).

R. Harré and E. H. Madden (1975) *Causal Powers: A Theory of Natural Necessity* (Oxford: Blackwell).

Marx, K. (1977 [1852]) 'The Eighteenth Brumaire of Louis Bonaparte' in D. McLellan (ed.) *Karl Marx: Selected Writings* (Oxford: Oxford University Press).

G. E. Myers (1986) 'Introspection and Self-Knowledge', *American Philosophical Quarterly*, 23(2), 199–207.

B. Stock (2010) *Augustine's Inner Dialogue: The Philosophical Soliloquy in Late Antiquity* (Cambridge: Cambridge University Press).

10

Reducing the Effort in Effortful Control

Stuart G. Shanker and Devin M. Casenhiser

In the fourth Book of *The Republic*, Socrates tells the following story:

> I once heard something that I believe, that Leontius, the son of Aglaion, was coming up from the Piraeus under the north wall from outside and observed corpses beside the public executioner. At the same time he had an appetite to look and again felt disquiet and turned himself away. For a while he fought and covered his face. But overcome by appetite (ὑπὸ τῆς ἐπιθυμίας), he stretched his eyes, ran towards the corpses and said, 'See for yourselves, you wretches, replenish yourselves with the beautiful sight.' (Rep. IV, 439)

Leontius, like all educated men, knew that he should not look, yet his impulse was overpowering. For Plato, there was nothing shocking about the fact that he should have had such a perverse urge; we all have such urges. Nor is it surprising that he should have been unable to control himself; we all get overpowered in this way. Such urges and internal conflicts are, Plato felt, a core feature of human nature.

Plato believed that we are at constant war with our impulses: that there is what he describes as a 'civil war' between two different parts of our mind: our faculty of reason, and our appetitive self. For the faculty of reason to acquire control over the appetites, it is imperative that it should be *educated*, so that we learn which are the 'right' sorts of desires and which are self-destructive. Yet, Leontius was clearly aware that he should not look. So, the reason why Socrates tells this story is to make the point that reason alone is not enough to vanquish the appetites: that it is up to a third faculty in the mind, that of *thumos*, or 'spiritedness', to see that reason triumphs (Rep. IV, 440).

Thumos, Plato argued, must be trained from an early point in a child's life to 'obey reason and be its ally' and to 'fight' against the appetitive self 'with courage' (Rep. IV, 442). If an individual's *thumos* is too spirited he will become, like Achilles in the *Iliad*, a 'wild beast' who cannot control his thoughts and actions (Shanker, 2012); but if it is not spirited enough the individual will, like Leontius, be unable to overcome his appetites. By no means is Plato faulting the appetites themselves; this is a fact of life, the urges and temptations that boil up from our animal nature. Hence, there is no point in railing against this elemental part of our constitution, and in so doing, Leontius reinforces the overall picture of character weakness.

If the faculty of *thumos* is not strong enough to carry out reason's dictates, the result will be 'injustice, licentiousness, cowardice, ignorance, and, in a word, the whole of vice' (Rep. IV, 444). It was, in fact, Plato who first insisted that if we lose the battle against our emotions and appetites, the result will be *mental* and not just physical illness; for 'to produce health is to establish a natural relation of control and being controlled, one by another' (Rep. IV, 444).

It is imperative, then, that reason should be cultivated, and spiritedness should be strengthened. How this is done must be tailored to the needs of the individual child. Plato distinguishes, for example, between the effects of too much flute music on a child with a weak *thumos* and its effect on one who has been born with a highly 'spirited nature' (Rep. IV, 411b). The former 'will become weak and dissolute', the latter 'quick-tempered, prone to anger and filled with discontent, rather than [becoming] spirited'. Overlooking what must strike us today as Plato's rather curious hostility to the flute, the important point here is Plato's insistence that a child's caregiving experiences should be tailored to suit the child's *thumos* (perhaps best likened to temperament today): something that Plato, long before Mary Rothbart sharpened our focus (Rothbart, 2011), saw in biological terms.

The story of Leontius has resonated with Western thinkers down through the ages.[1] Indeed, the concept of *thumos* not only influenced, but was likely the source of, our concept of *willpower* (Sorabji, 2002). To be sure, *thumos* had none of the moral connotations that are so strongly tied to the Early Christian view of willpower; yet, the image of the internal strength needing to vanquish the appetites has never wavered. Indeed, Plato's emphasis on the *mental effort* that an individual must make to inhibit his impulses, and the importance of cultivating a child's desire and ability to make such a mental effort, has served as one of the fundamental mainstays in western views of childrearing.

10.1 The Early Christian view

The story of Leontius represents the climax of a philosophical inquiry, beginning at *Republic IV 431* into the nature of self-control and how it is acquired. Socrates concludes that self-control is a function of two factors: education – viz, acquiring a set of 'rational' desires, and the development of *thumos*, so that the individual possesses the internal 'strength' to act on these desires and 'master' the 'irrational appetitive part' of the soul that craves 'certain indulgences and pleasures' (Rep. IV, 439).

From Plato onwards, it has been treated as a given that self-control is needed to prevent our appetites from interfering with our rational desires.[2] This theme was picked up at Galatians 5:17, when St. Paul warns: 'The desires of the flesh are against the Spirit, and the desires of the Spirit are against the flesh, for these are opposed to each other, to keep you from doing the things you want to do.'

But Early Christian thinkers introduced an important modification to this classical argument. They believed that, prior to the Fall, Adam and Eve enjoyed perfect humeral balance, with reason in full control of their desires and emotions, and thus, immunity from disease. By succumbing to the Devil's temptation, two critical things happened: first, our reason weakened, and second, our appetites strengthened (Midelfort, 2000). Hence, the need for self-control and the difficulty involved in exercising it were the direct consequence of the Fall and closely associated with Original Sin.

Plato's writings on self-control were largely concerned with the fear of death, but for the Early Christian thinkers, death is only to be feared if one has not struggled to control one's passions during one's life. In direct opposition to classical thought, Saints Gregory and Cassian highlighted the need to control anger, and in addition, add avarice, envy, gluttony, lust, pride and sloth to the list. These, of course, constituted the Seven Deadly Sins, a term which reinforces the most important aspect of the shift from classical to Early Christian thought: namely, that lack of self-control was not simply a matter of individual weakness, but the consequence of Original Sin.

Whether Early Christian thinkers intended the story of the Fall to be read literally or allegorically, their message was that it is a constant struggle for humans to control their impulses, which were thought to have a corrupting influence on the humours and thus to lead to an overheating of the brain and illness, both physical and mental (Arikha, 2008). Physical afflictions like deafness or blindness, and mental afflictions like madness or imbecility were regarded as the consequence of

failing to control one's impulses, and the individual was accordingly treated for his ignorance and lack of willpower.[3]

The resulting view of childrearing was that, for their own physical and mental well-being, impressionable young minds have to be *taught* to distinguish between True Desires (those that would lead to the fulfilment of God's Will) and false desires (those planted in us by Satan); and *disciplined*, so that the child acquires the internal strength to act on these True Desires. Herein lay the upshot of the doctrine of Original Sin: namely, that even – or perhaps, especially – young children are the repositories of sin (Shanker, 2008).

Accordingly, the twin goals of the medieval Quadrivium were to cultivate reason and strengthen the will so that the child would have the desire and would make the necessary effort to control his sinful nature. This view is captured by the famous carving at the western portal of Chartres Cathedral, which depicts Grammar as an older woman with a stern look on her face, sitting with a watchful eye over two young pupils. In her left hand, she holds an open book, and in her right, a flagellum with which she is about to beat one of the children who is misbehaving (Shanker, 2008). The fact that it was chosen to adorn a prominent religious structure underscores the association of the lack of willpower with sin – perhaps serving as a visible admonishment to young children. In any case, the image reinforces the notion that a child must have both knowledge and discipline drilled into him, even beaten into him if necessary.

This view of the Will as a faculty needing to be strengthened dates back to Plato's metaphor of appetites and emotions as 'wild horses' that need to be restrained: the stronger the animal, the stronger needs to be the hand wielding these reins. And one 'strengthens' this faculty in a child with 'mental exercises': a 'gymnasium for the mind' (Cribiore, 2001). The latter idea owes much to the ancient view of Memory as a kind of 'muscle' that could be 'strengthened' through appropriate exercises (Carruthers, 2008; Yates, 2010). According to this metaphor, the ability to recall long fragments of poetry was comparable to the ability to lift a heavy weight: One might be born with a natural aptitude, but regardless of this biological inheritance, one would nonetheless have to train just as assiduously to become an epic poet as to become an Olympic champion.

Willpower was thought to develop in much the same way that memory, indeed, that any muscle develops. For example, one raises a warrior by giving a young boy a sword and instructing him to hold it out at arm's length. If repeated over and over then, by the time he is

a young man, he will be able to hold the sword at arm's length with ease. So, too, a child must be trained to resist his impulses. And here is the critical point: Some children will require much more training than others, perhaps because their 'muscle' is just naturally weak, or perhaps because their impulses are unnaturally strong. For these children, the hand wielding the flagellum must not flag.

10.2 The act of inhibition

The accepted standard definition of self-control (i.e., willpower) today, or 'effortful control' as it is now commonly referred to in the scientific literature, is that it consists in 'the ability to inhibit a dominant response to perform a subdominant response' (Rothbart, 1989). The reason for this behaviouristic language is because of the desire to define effortful control in such a way that it is about inhibiting impulses *simpliciter:* as much a matter of attentional control as emotion-regulation and behavioural control, all subsumed under the rubric of *effortful control*.

For the moment, we need only note that where for the Ancient Greeks self-control was essentially about mastering the appetitive and emotional self, for moderns it is as much about controlling the sorts of attentional faculties required to perform a Stroop task (saying the *name* of the colour we are looking at rather than reading the colour-word that is written) as it is about resisting the impulse to grab a toy from another child or telling someone we have not seen for some time how shocked we are to see how much weight they have gained (Diamond, 2000). Indeed, for modern theorists, these different aspects of effortful control are closely intertwined; thus, for example, it has been shown that an early deficit in focused attention predicts later problems in behavioural control (Kochanska, Murray, and Harlan, 2000).

Measures of effortful control often include indices of attentional regulation (e.g. the ability to voluntarily focus or shift attention as needed, called 'attentional control') and/or 'behavioural regulation' (Eisenberg, Smith, Sadovsky, and Spinrad, 2004). The problem with this, however, remains the same as has historically plagued this issue: viz., the more we probe the various terms presented here, the more the definition leaves us, either with a different set of problematic terms, or else starts to look tautologous. For what exactly is it to *inhibit* a dominant response if not to make an effort to control it? To add the idea that we do so *wilfully* or *voluntarily* is as old as Origen and as puzzling as it was for the Church Fathers.

The core idea here remains the basic premise that effortful control involves *trying* to suppress a dominant response. But how exactly does one do this? And if a child is impulsive, does that mean that he did not try hard enough? Or that the impulses he was attempting to inhibit were simply too powerful for his meagre resources, for whatever reason? And does this entail, as the recent longitudinal studies on the long-term impact of poor delay-of-gratification at age four seem to indicate, that such a child is faced with a future of both mental and physical compromise?

Over forty years ago, Mischel developed his famous 'marshmallow test' (Mischel, Shoda, and Peake, 1988). In this test, a preschooler is seated at a table on which the experimenter places a marshmallow. The child is then told that she can eat the marshmallow now, or if she can wait until the experimenter returns, she will be given two marshmallows to eat. The experimenter then leaves the room for fifteen minutes and observes the child's behaviours as she waits for the experimenter to return. Results from this test consistently show that only around 30 per cent of four-year-olds can wait. But what is really striking about this test is that it turns out that the children who can wait, score an average of 210 points better on their college-entrance exams. And it is not just academic achievement that is at stake here: poor performance on delay-of-gratification tests predicts things like anti-social behaviour, internalizing problems, educational attainment, and susceptibility to drugs (Mischel, Shoda, and Peake, 1988).

The challenge here is to understand why a delay-of-gratification test can make such powerful predictions, for clearly the test is tapping into something important vis-à- vis a child's subsequent trajectory. Is it because their 'self-control muscle' tires more quickly? As Roy Baumeister has shown in a number of elegant experiments, tasks requiring sustained concentration, or emotional or behavioural control have a significant impact on a child's performance on a subsequent effortful-control task, and much more so in some children than others (Baumeister and Tierney, 2012). Is this because of some congenital neural deficiency?

10.3 The shift from mind to brain

For Plato, the mind (*psyche*) was the staging ground for the 'civil war' described above between reason and the appetites. Although Plato's view of the tripartite mind quickly faded, his idea of the mental effort required to win this internal conflict remained a constant theme over the next two millennia. But over the past decade, a fundamental shift

has occurred as scientists begin to study the neural processes that are thought to underpin 'effortful control'.

It is tempting, on the classical outlook, to see the shift that is going on today as simply involving the locus of the 'internal conflict'. That is, whereas for Plato it was the mind, for the modern developmental scientist it is the brain, with the neuroaxis serving as the staging ground for an internal conflict between the subcortex (brain stem and limbic system), which is the source of impulses and powerful negative emotions, and the prefrontal cortex, whose job it is to tame these powerful urges (e.g. by reappraisal, self-distraction, goal switching).

Roughly speaking, the *neuroaxis* is conceptualized as proceeding from the lowest or most primitive level of the brain (the brain stem) to the most advanced (i.e., phylogenetically newest) structures in the cerebral cortex. The oldest levels are the most structured at birth, while those at the upper end are the highly plastic structures that are fundamentally shaped by the child's experiences (Lewis, 2005; Lewis and Todd, 2007; Tucker, 2001).

According to recent research, the 'basic emotions' are triggered by a selective range of stimuli and set off a wave of physiological, behavioural and experiential responses. These basic emotions (e.g. love, anger, fear, curiosity) flow up the neuroaxis: that is, they are highly structured 'affect programs' (Ekman, 1992) that were formed in our prehistory. Not only is there a downward flow of control but also, an upward flow (e.g. from brain stem and hypothalamus) of synaptic activation and neurochemical stimulation. In so-called vertical integration, stimuli trigger upward flow (primitive physiological and emotional responses), which in turn are controlled by top-down processes. The bottom-up processes may enhance as well as direct top-down until the latter reduce emotional activation. This coordination may be marked by phase synchrony between 'higher' and 'lower' systems (Lewis, 2005; Tucker, 2001).

The crux of the neuroaxis hypothesis in regards to self-control is that the more time there is between stimulus and response, the more opportunity to select the most beneficial action: that is, the better the child's ability to stop and reflect, to choose between different goals, or to monitor and correct goal progress. But for some children, there is next to no pause between stimulus and response: no time whatsoever to choose between goals, and hence, no opportunity to exercise self-control.

Such children are going to have to make a much greater effort to inhibit their impulses. The cause of their problem might be a limbic system that is particularly 'arousable' or 'reactive'; or the problem might lie in a

medial prefrontal cortex that is somewhat limited for genetic reasons, or has been weakened by smoking or alcohol during pregnancy, and possibly even by maternal stress (Fogel, 2009; Sagvolden et al., 2005); or perhaps the child's problem is a shortage of those neurohormones that help a child negotiate between short-term desires and long-term rewards (Sagvolden et al., 2005). But whether it is because the impulses are stronger or the dampening systems are weaker, these children are thought to have much more difficulty with inhibition.

As can be seen from the foregoing discussion, as modern as this argument might sound, the basic picture operating here dates back to Plato. Take, for example, Libet's famous experiments (Libet, 2004), which were widely thought at the time to show that in voluntary actions, an electrical signal (the 'readiness potential') precedes the action by around 200 milliseconds.[4] To be precise, Libet claimed that these experiments show that we have a very small window of time in which to consciously suppress an unconscious impulse (Libet, 2004). As Ramachandran pointed out, what Libet was really suggesting is that the theory of 'free will' should be recast as a 'theory of free won't' (Ramachandran, 2011). And that, of course, is very much the point that Socrates was making with the story of Leontius.

To be sure, the modern theorist has recourse to the tools afforded by developmental neuroscience to study what is happening in a child's brain when he is engaged in things like a frustrating go/no-go task. For example, we can measure the amplitude of frontal N2, the ERP that occurs around 200–400 milliseconds after a stimulus, and the error-related negative N2 and ERN are still widely interpreted as reflecting the effort involved in inhibiting a response (Falkenstein, Hoormann, and Hohnsbein, 1999; Jodo and Kayama, 1992).[5] But it is a mistake to think of the brain as exerting an effort to produce these markers. The brain, or part of the brain, does not make an effort: it simply processes information. To be sure, some processes require greater amounts of energy as we explain below (see also Porges, 2011), but this is not to be confused with effort.

These studies which investigate what is happening in the brain when a child is struggling with a frustrating or an emotionally challenging task provide an important insight into why some children find it so difficult to stay calmly focused and alert. What these studies *cannot* explain, however, is the source of the child's effort, any more than could Plato's *thumoeidic psyche* or St. Augustine's *plena voluntas*; for the question of in what part of the mind or the brain the effort in effortful control resides is metaphysical, not empirical, and one key to understanding the

significance of contemporary neuroscientific findings is to understand the significance of this fundamental philosophical distinction.

10.3.1 The metaphysical problem of self-control

Throughout his writings, beginning with the *Tractatus* and right up to his *Last Writings on the Philosophy of Psychology*, Wittgenstein was fascinated by the problem of the will. His basic idea, encapsulated in a brief remark in *Philosophical Investigations,* was that:

> I can't will willing; that is, it makes no sense to speak of willing willing. 'Willing' is not the name of an action; and so not the name of any voluntary action either. And my use of a wrong expression came from our wanting to think of willing as an immediate non-causal bringing-about. (PI, §613)

What Wittgenstein is warning against here is the assumption that when we do something wilfully, it must be because we first experience a mental act, a *volition*, that causes us to behave (or not to behave) in a certain way. This assumption then sets us off on the search to discover the workings of this mysterious 'volitional' part of the mind. And if, like Leontius, we fail to 'inhibit a dominant response to perform a subdominant response', it must be because our 'volition' simply wasn't up to the task. And the key to escaping from this metaphysical trap is to recognize that 'willing' is not the name of a mental action or experience.

As always when reading Wittgenstein, one might be left more puzzled by his enigmatic remarks than the problem we are supposed to be addressing. But his *Nachlass* clarifies the thinking behind this argument. In *Remarks on the Philosophy of Psychology* Wittgenstein asks: 'How is "will" actually used' (RPP I, §51). The reason why this question is so important is because: 'In philosophy one is unaware of having invented a quite new use of the word. ... It is interesting that one constructs certain uses of words specially for philosophy, wanting to claim a more elaborated use than they have.' As he made clear in *Culture and Value* (p. 15), it was the ancient Greeks who inspired the problematic uses of 'will', 'effort', 'self-control' that continue to frustrate our efforts to understand effortful control; for they treated self-control as the name of a mental act that, depending on the amount of mental effort exerted, succeeds or fails to inhibit an impulse.

To be sure, there are all sorts of ordinary, non-problematic uses of volitional terms: '"Want"', for example, 'is sometimes used with the meaning "try": "I wanted to get up, but was too weak"' (RPP I, §51). The

metaphysical problem only arises when one assumes that: 'Wherever a voluntary movement is made, there is volition. Thus if I walk, speak, eat, etc., then I am supposed to will to do so. And here it can't mean trying' (RPP I, §51). But that is absurd; for there are only special circumstances in which we speak of *trying* to do something: 'when I walk, that doesn't mean that I try to walk and it succeeds' (RPP I, §51). We only speak of *trying to walk* when there is some reason to indicate that walking would be difficult; if, for example, it is very windy, or I am recovering from an illness. And we only speak of *trying to inhibit an impulse* when there is something about the situation that licenses this description.

That does not mean that there is not such a thing as mental effort; but here, too, one has to consider:

> How do I learn [that] the words 'I exert myself', 'This is hard', 'Ugghh' when lifting a weight is the natural expression. We can all imitate a man lifting a weight. If a child grunted when it turned over pages and not when it lifted a weight, this would be confusing. (LPP, p. 36)

That is, what are the circumstances in which we learn how to use expressions like 'mental effort'? To be sure, we draw on the language of 'physical effort' when talking about the mental effort involved in working (*sic!*) on a problem. We speak of *wrestling with the problem*, of being left *exhausted after our mental exertions*. But then, these are classic examples of 'metaphors that we live by' (Lakoff and Johnson, 1980) – metaphors that shape the way we think about and indeed experience 'mental effort'.

Moreover, we certainly understand what it means to say: 'enter not into temptation: the spirit indeed is willing, but the flesh is weak' (Matthew 26: 41). But the point of Wittgenstein's philosophical critique is not that there is no effort involved in willpower; rather, it is that, as Peter Hacker explains:

> [Willpower is not a] mental correlate of muscle power, and strength of will is not a matter of having causally efficacious volitions. It is rather determination, persistence, and tenacity in pursuit of one's goals. But it is a philosophical fiction to suppose that our voluntary actions are always preceded by acts of will, or that the power of the will is exhibited in ordinary action as the power of the muscles is exhibited in ordinary movement. (Hacker, 2000, p. 583)

To say that willpower is not a correlate of muscle power is not to deny that it makes sense to speak of the strength of one's willpower; but this

is to describe how a *person* responds to a challenge or a temptation, not the amplitude of a neural waveform when they are so doing. For it is the person who makes this effort, not his brain; brains do not make any effort to process stimuli, they just process them. And the source of that individual's strength is located, not in his prefrontal cortex, but rather, in his motivation or his beliefs or his upbringing.

Moreover, one would only use this term, according to Grice's maxims, if there were some particular feature of the action to which one wanted to draw attention; for, to paraphrase Wittgenstein, it does not (ordinarily!) require willpower to say, 'Good morning' to a neighbour. And to say that 'S just didn't have enough willpower to stay on his diet over Xmas' is not to say that some mechanism in his mind or brain was not strong enough to control another part of his mind or brain (e.g. his desire, or his limbic system), but that S gorged himself on shortbread knowing full well that he would pay for this act of indulgence in the new year.

10.3 2 The story of Leontius seen through a Wittgensteinian lens

When Leontius claims that he could not control himself because he was 'overcome by desire' *(*ὑπὸ τῆς ἐπιθυμίας*),* he was not *describing* but rather was *expressing* his mental conflict. That is, the first-person statement 'I could not control myself because I was overcome by desire' is not an empirical description which, by definition, must be capable of being either true or false but rather, is an *avowal*; that is, a unique kind of non-contingently true first-person utterance in which a speaker gives a reason for his action.[6] Leontius does not possess privileged access that enables him to observe and relate the inner workings of his mind; rather, his statement operates as a criterion for our saying that 'The reason why Leontius looked was because he was overcome by his macabre desire to see the corpses' (Hacker, 2000, see ch. 2).

Were one to treat this avowal as an empirical proposition, then what else could Leontius have been describing when he said, 'I could not control myself' other than, for example, the fact that his 'faculty of *thumos*' (or 'willpower') was not strong enough to execute the dictates of his reason? The obvious concrete analogy is that someone could have forcibly prevented Leontius from looking at the corpses, and if Leontius had been a big strapping fellow, it would have taken a great deal of effort to hold him back. This physical struggle might be taken as analogous to the internal struggle being waged between the different parts of Leontius' mind (or brain). So, Socrates reasons metaphysically that what a story like that of Leontius must be telling us is that the part of his mind that performed this regulatory function simply was not up to the task:

it was not strong enough to inhibit the impulse. But what Wittgenstein shows us is that words like 'strength of will' describe how an *individual* deals with temptation: not something that was happening inside his mind when he acted.

In fact, it seems clear that Leontius is struggling with two opposing desires: one to satisfy his morbid curiosity, and the other to satisfy his desire to maintain proper decorum and dignity. What happens is that Leontius *chooses* one desire over the other, and in this case it is a shameful, or at the very least what he himself regards as the improper, desire that he chooses to satisfy. But rather than publicly announce, 'I chose to do something of which I am ashamed', Leontius attempts to shift the blame to the *desire itself*, to his appetitive self, a force so powerful that it could not be contained.

The critical point here is not the insight that psychological defence mechanisms are reflected in the creation of metaphors (though this is interesting), but that the concept of self-control is internally tied to an individual's capacity to inhibit an impulse (Strawson, 2008). If we judge that Leontius was indeed weak, it is not because of something that may or may not have been going on in his mind (or his brain), but because we feel that he *could* have acted otherwise but failed to make a strong enough effort to do so. But perhaps there were mitigating factors?

It was Hume who recognized the significance of this point and questioned whether there might be hidden dimensions to this story that Plato's audience would have recognized but that are lost on us (Paton, 1973). Would the fact that Plato so carefully specified which Leontius he was talking about have served to indicate the reason why Leontius experienced this emotionally-charged reaction?[7] Or perhaps Leontius was ill, or had been on a long journey, which left him exhausted? Would he have still have chosen to look at the corpses if he had been healthy and well rested when he came upon the site of the execution?

It is a particularly apt question when we attempt to unravel the significance of something like a delay-of-gratification test performed on a four-year-old child. The upshot of Wittgenstein's argument is that we need to *look at the child* before we talk about him as 'trying' or 'failing' to 'inhibit an impulse', let alone describing a child as 'weak'. For example, we would never get angry at a child with autism for having a tantrum, and we would never tell his parents that their child needed a dose of 'tough love'. But why is the four-year-old who snatches up the marshmallow any different? That is not to say that the task is without significance; the challenge is, rather, to clarify where that significance lies.

10.3.3 The importance of self-regulation

Every stress that a child must face demands energy. The stressors in question might be physiological, emotional, cognitive, social, or prosocial (Shanker, 2012). The cost to the child's autonomic nervous system will vary according to the stressor and, of course, the child's physical or emotional state. Two children might have to expend very different amounts of energy to deal with the same situation. Suppose, for example, that we are dealing with a child who finds sitting in a classroom very stressful, for different reasons. Perhaps he finds the visual and auditory stimuli distracting, and he has to work hard to filter this out in order to pay attention to his teacher; or he finds the hard seat uncomfortable, and it is taxing for him to sit still for too long. In cases such as these, the child is not only expending a great deal of energy attending to his teacher, but considerable amounts of energy trying to inhibit the distraction or to counteract the forces of gravity (Porges, 2011).

Of all the costs on a child's energy reserves, one of the greatest is anxiety. Anxiety sets off a vicious cycle in which the intensity of an impulse is heightened, demanding yet further energy to inhibit that impulse. What the research on N2 is showing us is that some children become so much more anxious than others on a frustrating or emotionally challenging task. There are any number of reasons why this might be the case. It might relate to the child's biological constitution, parenting practices, or the medium of the task itself (Shanker, 2012).

It is no surprise, given the tight interconnection between arousal and attention, that the more anxious a child, the more constrained his attention span, and the more likely that the child will tend towards either a hypo-aroused state in which he shuts down to try to restore energy, or a hyper-aroused state in which his impulsivity is heightened. In either case, what is needed is not a greater effort, either to 'pay attention' or to 'control his impulses'; rather, what is needed, especially with a young child, is a reduction in the stressors that are taxing and draining his system, leaving him in such a vulnerable state.

This is precisely the reason, we believe, why delay-of-gratification tests performed on a four-year-old can have such long-term physical and mental health implications (Moffitt et al., 2011). For what these tests may be showing us, at least in part, is that some four-year-olds are experiencing far too many stressors (Mustard, McCain, and Shanker, 2007). If the sources of these stressors are not addressed, and the consequent load on the child's autonomic nervous system reduced, then we see the sorts of downstream psychological and behavioural

consequences that have been noted in the 'ego depletion' literature (Shanker, 2012).

The problem is not that some children are *weaker*, therefore, but that some children have to work much harder than others to perform the same tasks, and it is this expenditure that so seriously depletes their capacity to meet subsequent challenges. This, we suspect, is a key reason why so many children find the marshmallow task difficult. And this is the reason why the task has the predictive significance that it does in regards to the child's long-term physical and mental well-being.

A child who daydreams excessively or is inordinately hyperactive is certainly not culpable in any way, and it would be deeply unfortunate to treat the child as if he were, however unconscious this might be. Rather than trying to *strengthen* their ability to remain focused and alert (e.g. through punishment and reward), we need to understand and thereby mitigate the drains on their nervous system that have resulted in their chronically depleted physiological state.

10.4 The MEHRIT study

For the past seven years, we have been studying the effectiveness of a parent-mediated intervention for young children with autism (MEHRIT) based on Greenspan and Wieder's DIR (Individual Differences/Relational model) (2006). Our basic assumption was that the reason why children with autism typically have so much trouble in social interactions is not because they are lacking a fundamental social need (viz., the 'belongingness drive'; Baumeister and Leary, 1995); nor because they are congenitally incapable of interacting socially (e.g. because they are born with a defective 'theory of mind mechanism'; Baron-Cohen, 1995). Rather, the reason why these children have so much trouble with social interactions is because of the amount of stress involved.

The source of their stress might be a sensory system that is overloaded by various aspects of the interaction: for example, the myriad and rapidly changing signals involved, the noise, visual stimulation, even the odours. The child with autism's typical responses – for example, gaze aversion, self-stimulating, perseveration – represent defensive behaviours that the child adopts to deal with an encounter that he finds overwhelming. The problem is that the child's mode of self-regulating blocks the development of social skills, and indeed, of the social brain network, thereby exacerbating the stresses involved in social interaction (Shanker, 2012).

Our goal was to understand and thereby reduce the stresses the child experiences so as to enhance his desire and ability to engage in social interactions. What we found was that MEHRIT has a dramatic impact on the pleasure that the child experiences in social interactions; his desire to initiate social interactions; his ability to remain engaged in social interactions; and his development of the communicative skills that are needed to be so engaged (Casenhiser, Shanker, and Stieben, 2013). But the larger lesson that we took away from this study is that we needed to write a paper on 'reducing the effort in effortful control'; for what we came to recognize is that far too many children are dealing with far too much stress in their lives, because of biological, social, psychological, and/or environmental reasons.

Quite simply, these children have to work much harder to stay calmly focused and alert, and an allostatic load condition will have ever greater downstream effects (e.g. on language development, social development, mood, impulsivity) as the negative effects of poor self-regulation lead them to fall further and further behind their peers, or have greater and greater social, psychological or health problems, thereby exacerbating the drain on their already over-stretched system.

The first and most critical step to helping these children is to reframe their behaviour, and to accomplish this we must leave behind the classical and medieval views of self-control. As we have seen, these views assume that self-control is simply a matter of exerting the effort required to inhibit an impulse, and thus, that a child who is not exerting a sufficient effort in this regard is somehow being weak or headstrong and should be treated accordingly. But for most children, punitive 'corrective' actions actually exacerbate their problems with self-control (O'Keefe, 2005). This is because such actions can add to the excessive stress load that they are already dealing with.

Rather than becoming angry or irritated by a child's lack of self-control, we should always assume that a child *would* exercise self-control *if* he could. To be sure, this is an enormously bold statement. It is not just a matter of questioning the effectiveness of some of our disciplinary practices. Rather, it involves a fundamental and far-reaching shift in the manner in which we look at children.

Instead of assuming that a child is being wilfully oppositional or inattentive, our assumption should always be that children do not like disappointing the important adults, and for that matter, peers in their lives: that they themselves may not understand why they lashed out or were confrontational or even why they are constantly getting in trouble. And if that is the case, we need to look past their behaviour in order to get to the heart of what is really troubling them.

The implications of the view of effortful control that we have sketched in this chapter are clear: It is not just antiquated, but counter-productive, to blame a young child for whatever problems he might be experiencing in emotional, behavioural, or attentional control. If we want to prepare that child for the world of social learning that awaits him, then it is imperative that we understand the sources of his difficulties. But there is a deeper point here.

The title of this chapter questions the persisting influence of ancient Greek attitudes in our modern views of effortful control; but in no way do we seek to question the importance of effortful control. And neither, although we speak of reducing the effort in effortful control, is our intention to question the importance of effort. But to the extent that perseverance is fuelled by success, effort begets effort.

What we have tried to do at MEHRIT is change the trajectory for the children that we work with: to open them up to a world of social learning, in all its many forms, that is only made possible by the pleasure they experience with their caregivers and the effort they make to seek out these experiences. But the more effort they have to expend trying to stay calmly focused in an environment that they find overwhelming, the less they will be able acquire the self-regulating skills touched on in this chapter.

Notes

1. This research was made possible by the generous support of the Harris Steel Foundation and the Harris family, which made it possible to create the Milton and Ethel Harris Research Initiative (www.mehri.ca). We are also grateful for the support we have received from the Unicorn Foundation, Cure Autism Now, the Public Health Agency of Canada, the Templeton Foundation, and York University. We would like to record our deep debt to the clinical team at MEHRIT collectively known as FACE: Fay McGill, Amanda Binns, Chris Robinson and Eunice Lee.To learn more about this fascinating history, see the essay on 'The Will' that Hans Oberdieck wrote for Harry Parkinson's *Encyclopaedia of Philosophy* (1988).
2. Not to mention behaviours that are harmful to others, a theme that was highlight by the ancient Greeks. For example, Medea confesses (in the play of the same name by Euripides): 'I am overcome by evil, and I realize what evil I am about to do, but my passion controls my plans.'
3. Recall that it was Plato who recommended that the State establish *sophronesterion*: correctional institutes where the insane would have their madness drilled out of them by 'instruction' (Laws X).
4. There has been considerable debate about both the methodology and the interpretation of Libet's experiments; for probing analyses of the conceptual problems involved in this reading of Libet's experiments, see (Bennett and Hacker, 2003; Coulter, 1989).

5. But see (Nieuwenhuis et al., 2003), who argue that the data are also consistent with the interpretation that N2 is simply an indicator of the detection of response conflict.
6. The fact that Leontius gives this as his reason does not, of course, entail that this actually does explain his behaviour. As Plato so brilliantly explored in his analysis of Achilles, such an explanation might look at the kind of upbringing that Leontius experienced and why this failed to help him develop the self-control that he clearly desired in this situation (Shanker, 2012).
7. Were they, Hume asked, members of the 'Thirty Tyrants' who were overthrown in 403 BCE, some of whom were relatives of Plato? And was the point of the story that Leontius was deeply conflicted about their execution: torn, e.g. between personal ties and his feelings about democracy?

References

N. Arikha (2008) *Passions and Tempers: A History of the Humours* (New York: Harper Perennial).

S. Baron-Cohen (1995) *Mindblindness: An Essay on Autism and Theory of Mind* (Cambridge, MA: MIT Press).

R. Baumeister and J. Tierney (2012) *Willpower: Rediscovering the Greatest Human Strength* (New York: Penguin Group USA).

R. F. Baumeister and M. R. Leary (1995) 'The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation', *Psychological Bulletin*, 117(3), 497–529.

M. R. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Malden, MA: Wiley-Blackwell).

M. J. Carruthers (2008) *The Book of Memory: A Study of Memory in Medieval Culture* (New York: Cambridge University Press).

D. M. Casenhiser, S. G. Shanker, and J. Stieben (2013) 'Learning through Interaction in Children with Autism: Preliminary Data from a Social-Communication-Based Intervention', *Autism*, 17(2), 220–41.

J. Coulter (1989) *Mind in Action* (Atlantic Highlands, NJ: Humanities Press International).

R. Cribiore (2001) *Gymnastics of the Mind: Greek Education in Hellenistic and Roman Egypt* (Princeton, NJ: Princeton University Press).

A. Diamond (2000) 'Close Interrelation of Motor Development and Cognitive Development and of the Cerebellum and Prefrontal Cortex', *Child Development*, 71(1), 44–56.

N. Eisenberg, C. L. Smith, A. Sadovsky, and T. L. Spinrad (2004) 'Effortful Control: Relations with Emotion Regulation, Adjustment, and Socialization in Childhood' in R. Baumeister, K. Vohs (eds) *Handbook of Self-Regulation: Research, Theory, and Applications* (New York: Guilford Press), 259–82.

P. Ekman (1992) 'Are There Basic Emotions?' *Psychological Review*, 99(3), 550–2.

M. Falkenstein, J. Hoormann, and J. Hohnsbein (1999) 'ERP Components in Go/No-go Tasks and Their Relation to Inhibition', *Acta Psychologica*, 101(2–3), 267–91.

A. Fogel (2009) *The Psychophysiology of Self-Awareness: Rediscovering the Lost Art of Body Sense* (New York: Norton).
S. I. Greenspan and S. Wieder (2006) *Engaging Autism: Using the Floortime Approach to Help Children Relate, Communicate, and Think* (Cambridge, MA: Da Capo Press).
P. M. S. Hacker (2000) *Wittgenstein: Mind and Will, Analytical Commentary on the Philosophical Investigations* (Malden, MA: Wiley-Blackwell).
E. Jodo and Y. Kayama (1992) 'Relation of a Negative ERP Component to Response Inhibition in a Go/No-go Task', *Electroencephalogr Clin Neurophysiol*, 82(6), 477–82.
G. Kochanska, K. T. Murray, and E. T. Harlan (2000) 'Effortful Control in Early Childhood: Continuity and Change, Antecedents, and Implications for Social Development', *Developmental Psychology*, 36(2), 220–32.
G. Lakoff and M. Johnson (1980) *Metaphors We Live By* (Chicago: University of Chicago Press).
M. D. Lewis (2005) 'Self-Organizing Individual Differences in Brain Development', *Developmental Review*, 25(3–4), 252–77.
M. D. Lewis and R. M. Todd (2007) 'The Self-Regulating Brain: Cortical-Subcortical Feedback and the Development of Intelligent Action', *Cognitive Development*, 22(4), 406–30.
B. Libet (2004) *Mind Time: The Temporal Factor in Consciousness* (Boston, MA: Harvard University Press).
H. C. E. Midelfort (2000) *A History of Madness in Sixteenth-Century Germany* (Stanford, CA: Stanford University Press).
W. Mischel, Y. Shoda, and P. K. Peake (1988) 'The Nature of Adolescent Competencies Predicted by Preschool Delay of Gratification', *Journal of Personality and Social Psychology*, 54(4), 687–96.
T. E. Moffitt, L. Arseneault, D. Belsky, N. Dickson, R. J. Hancox, H. Harrington, and A. Caspi (2011) 'A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety', *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), 2693–8.
J. F. Mustard, M. McCain, and S.G. Shanker (2007) *Early Years Study II* (The Council of Early Child Development: Toronto).
S. Nieuwenhuis, N. Yeung, W. van den Wildenberg, and K. R. Ridderinkhof (2003) 'Electrophysiological Correlates of Anterior Cingulate Function in a Go/No-Go Task: Effects of Response Conflict and Trial Type Frequency', *Cognitive Affective Behavioural Neuroscience*, 3(1), 17–26.
M. O'Keefe (2005) 'Teen Dating Violence: A Review of Risk Factors and Prevention Efforts', *National Electronic Network on Violence Against Women*, 1–13.
G. H. R. Parkinson and T. E. Burke (1988) *An Encyclopaedia of Philosophy* (New York: Routledge).
M. Paton (1973) 'Hume on Tragedy', *The British Journal of Aesthetics*, 13(2), 121.
S. W. Porges (2011) *The Polyvagal Theory: Neurophysiological Foundations of Emotions, Attachment, Communication, and Self-Regulation* (New York: Norton).
V. S. Ramachandran (2011) *The Tell-Tale Brain: Unlocking the Mystery of Human Nature* (New York: William Heinemann).
M. K. Rothbart (1989) 'Temperament and Development' in G. Kohnstamm, J. Bates, and M. K. Rothbart (eds) *Temperament in Childhood* (Chichester, UK: Wiley), 187–248.

—— (2011) *Becoming Who We Are: Temperament and Personality in Development* (New York: The Guilford Press).
T. Sagvolden, E. B. Johansen, H. Aase, and V. A. Russell (2005) 'A Dynamic Developmental Theory of Attention-Deficit/Hyperactivity Disorder (ADHD) Predominantly Hyperactive/Impulsive and Combined Subtypes', *Behavioural and Brain Sciences*, 28(3), 397–419.
S. G. Shanker (2008) 'In Search of the Pathways That Lead to Mentally Healthy Children', *Journal of Developmental Processes*, 3(1), 22–3.
—— (2012) *Calm, Alert and Learning: Classroom Strategies in Self-Regulation* (Toronto: Pearson).
R. Sorabji (2002) *Emotion and Peace of Mind: From Stoic Agitation to Christian Temptation* (Oxford: Oxford University Press).
P. F. Strawson (2008) *Freedom and Resentment and Other Essays* (New York: Routledge).
D. M. Tucker (2001) 'Motivated Anatomy: A Core-and-Shell Model of Corticolimbic Architecture', *Handbook of Neuropsychology*, 5, 125–60.
F. A. Yates (2010) *The Art of Memory* (New York: Routledge).

11
The Concepts of Suicidology

Michael D. Maraun

Suicidology is the branch of health science that is concerned with suicide, self-injurious behaviour, attempted suicide, and like phenomena. One focus of suicidologists is the formulation of explanations of these target phenomena, more particularly, explanations of *why* individuals commit suicide and perform other self-injurious acts. The explanations that have been formulated thus far have quite frequently involved the concepts *intention*, *motive*, and *reason* (Hjelmeland and Knizek, 1999), but, as Hjelmeland and Knizek have documented, these concepts are employed by suicidologists in equivocal, contradictory, and incoherent fashions. While conceptual confusion is commonplace within the social, behavioural, and health sciences, the suicidologists response to the conceptual confusion inherent to his line of investigation is not. For the suicidologist has both correctly identified certain of the problems that have impeded scientific progress within this line of investigation as *being* conceptual in nature and endeavoured to resolve these problems (see Hjelmeland and Knizek, 1999).

Unfortunately, the means by which the suicidologist has sought to resolve the conceptual confusions that undermine his empirical work are themselves confused. For although a resolution of conceptual confusions of the sorts that are of concern to the suicidologist can be achieved only through a careful analysis of the linguistic rules that fix the meanings of problematic concepts (cf. Baker and Hacker, 1982; Bennett and Hacker, 2003; Schulte, 1993; ter Hark, 1990), the suicidologist has turned for guidance to various of the causal theories of meaning that have arisen in psychology and philosophy.

These theories of meaning presuppose profound misconceptions about language; most particularly about the *nature* of concept meaning itself. In consequence, they manifest both confoundings of empirical,

conceptual, and metaphysical issues and a systematic *disregard* for the linguistic rules that must be clarified in a clarification of a concept's meaning. The suicidologists installing of these theories as the foundation for attempts to resolve the conceptual problems indigenous to this research area has, therefore, ensured lack of progress in that regard.

In this chapter, I will endeavour to provide clarifications of the concepts *intention*, *motive*, and *reason*, an understanding of the meanings of which has, by his or her own admission, eluded the suicidologist. Because these concepts have been especially problematic in their roles as ingredients of explanatory accounts of the phenomena investigated by suicidologists, these clarifications will, of necessity, have to be meshed with a satisfactory description of the phenomena themselves.

The organization of the chapter will be as follows: first, I will provide a (necessarily spartan) account of the conceptual problems endemic to the area; second, I will provide a description of the phenomena the explanatory accounts of which contain the problematic concepts; third, I will deal with the problematic concepts and, in particular, address the issue of their legitimate roles within explanatory accounts of these phenomena.

11.1 Conceptual problems

The concepts *intention*, *motive*, and *reason* frequently appear in the explanations that are formulated by suicidologists of phenomena of interest to them. Unfortunately, suicidologists employ these concepts in equivocal, contradictory, and incoherent fashions. As documented by Hjelmeland and Knizek (1999), this state of affairs is perhaps most salient in the tacit synonymizing of these and other concepts. Thus, Lukianowicz (1972) equates the concepts *aim* and *motive*, Bancroft, Skrimshire, and Simkin (1976), *reason* and *motive*, Birtchnell and Alarcon (1971), *intention* and *motivation* (whose relation to *motive* they do not explain), and Kovacs, Beck, and Weissman (1975), *reason*, *goal*, *desire*, and *motive*.

Against this backdrop of rampant synonymizing, there have been offered up a great many definitions, the justifications for which are unclear, and the agreement among which is, to say the least, lacking. The analysis of Hjelmeland and Knizek (1999) reveals suicidologists as frequently taking the synonymized *intention*, *motive*, and *reason* as standing, in a vague, unresolved fashion, for '…something the patient wanted to happen in the future, something they wanted to achieve by their act…' (p. 276). Whatever be the merits of this view, it does not square with either Alicke, Weigold, and Rogers (1990), in which it

is claimed that '... motives supply the reason for the desire', or Maris, Berman, and Silverman (2000, p. 37), on whose account a motive is '... the cause or reason that moves the will and induces action'.

On the one hand, Beck, Schuyler, and Herman (1974, p. 45) define suicidal intent as '... the *seriousness or intensity of the wish* [italics added] of a patient to terminate his life'. On the other hand, Maris, Berman, and Silverman (2000, p. 37) claim that an intention is '... the purpose a person has in using a particular means (e.g. suicide) to effect a result.' Whatever be the merits of *these* accounts, they square with neither the view of *intention* described by Hjelmeland and Knizek (1999) as being the standard within suicidology, nor the oft-cited treatment of Trevarthen (1982), in which intentions are described as being *causally brought about* by motives, before going on to *generate* intentional acts that satisfy these motives.

All told, perhaps the most forthright assessments come from the papers of Hjelmeland and Knizek (1999) and Velamoor and Cernovsky (1992), each of which contains a simple admission of uncertainty as to the meanings of *intention*, *motive*, and *reason*.

11.2 Conceptual clarifications

The role that a concept can legitimately play in an explanation of a particular explanandum is a function of: (1) the meaning of the concept (i.e., the linguistic rules that fix its correct employments); and (2) the nature of the explanandum. An explanation of *why* an individual performed an involuntary act σ, hiccoughing or some such, cannot legitimately involve a citation of the individual's *reasons* for having performed σ, because an individual *can have no reasons* for having performed an involuntary act. Causes are, instead, the legitimate ingredients of such explanations (Bennett and Hacker, 2003). An individual's *reasons* are, on the other hand, frequently the legitimate and relevant contents of an explanation of why he performed an act φ, in the case in which φ is an intentional act (e.g. the act of choosing a certain brand of shampoo).

The aim of the chapter is, indeed, to provide clarifications of the concepts *intention*, *motive*, and *reason* that suicidologists have identified as problematic. However, within the area of suicidology, these concepts are problematic in the context of their being ingredients of explanations of the phenomena that suicidologists investigate. This fact necessitates that I clarify the concepts that denote the phenomena that are the explananda of these explanations. I will focus my attentions on two explananda important to the area of suicidology: committing suicide

and attempting to commit suicide. The clarifications that I present in this section of the chapter are organized as follows: first, I deal with the explananda; second, with the problematic concepts *intention*, *motive*, and *reason*; and, third, I bring what has been learned to bear on the issue of the formulation of coherent explanations involving the concepts *intention*, *motive*, and *reason*.

11.2.1 Natures of the explananda

In the social, behavioural, and health sciences, the phenomena to be explained are very often the *doings* of humans. Committing suicide and attempting to commit suicide are doings. The task at hand, then, is to clarify the *kinds* of doings they are. To this end, I will depend heavily on the superb treatment of the topic of volition and voluntary movement that is found in Bennett and Hacker (2003) and a useful analysis of intentional acts by Kenny (1966).

Doings can be exhaustively classified as *acts* or *non-acts*. Acts are doings that humans perform (e.g. brushing teeth, picking up a phone, throwing a ball) or fail to perform (e.g. omissions, abstentions, refrainings) (Bennett and Hacker, 2003). Doings that are non-acts include such things as falling asleep, passing out, getting sick, dying, and tripping.

Acts can, in turn, be exhaustively classified as *voluntary* or *nonvoluntary*. Among other things, a voluntary act: (a) involves the exercise by the performing agent of a two-way power to perform or refrain from performing (this being the chief point of contrast between they and acts that are not voluntary); (b) involves the control by the performing agent of the act's inception, continuation, and termination; (c) can be engaged in at will; and (d) is an act that an agent can *try*, *intend*, or *decide* to perform (Bennett and Hacker, 2003).

Voluntary acts can be exhaustively classified as *intentional*, *unintentional*, or *neither intentional nor unintentional* and acts that are nonvoluntary, as *intentional*, *involuntary*, or *neither intentional nor involuntary* (Bennett and Hacker, 2003). Before fleshing out *these* distinctions, it will prove advantageous to review certain elements of Kenny's (1966) analysis of intentional acts. Following von Wright (1963), Kenny distinguishes between an *act a*, the *result a'* of act *a*, and a *consequence b'* of act *a*. The result of performing the act of closing the door is the door's being closed. One consequence of doing so may be that the cat cannot get into the room. The relation between an act *a* and its result *a'* is internal (cf. ter Hark, 1990, p. 47); that is, to be able to identify *a* is to be able to identify *a'*. The relation between an act *a* and one of its consequences *b'*, on the other hand, is empirical; *a'* brings about the state of affairs *b'*.

Let A be a human agent; a and b be acts; a' and b' be their results; $a' \rightarrow b'$ represent the state of affairs that b' is brought about by a'. On Kenny's analysis:

Ki) If Agent A performs a, he also performs b;
Kii) An act a performed by Agent A is an *intentional act* if A: (a) knowingly performs a; and (b) wants to perform a for its own sake or in order to perform b (1966, p. 647);
Kiii) An act b performed by Agent A is an intentional act if A: (a) knowingly performs a; (b) knows that in performing a he is performing b; and (c) wants to perform b (1966, p. 647);
Kiv) Agent A intends the result of any intentional act he performs;
Kv) Agent A foresees the result of any act he knowingly performs (hence, foresees the result of any intentional act he performs).

For analytical purposes, Kenny (1966) considers a number of scenarios in which are situated intentional acts. His scenarios 1, 2, and 9, which I will now review, would appear to be especially relevant to the discipline of suicidology:

> Scenario 1. Agent A knowingly performs a, wants to perform a, knows or believes that he is performing b by performing a, and wants to perform b.
>
> Scenario 2. Agent A knowingly performs a, knows or believes that he is performing b by performing a, and wants to perform b, but does not want to perform a for any other reason than to perform b.
>
> Scenario 9. Agent A knowingly performs a, wants to perform a, but does not know that he is performing b by performing a, and does not want to perform b.

On Kenny's analysis, then: (a) the act as of Scenarios 1, 2, and 9 are intentional; (b) the act bs of Scenarios 1 and 2, but not of Scenario 9, are intentional.

Let us now flesh out the category of voluntary acts:

Vi) A voluntary act a performed by an agent A is *intentional* if it satisfies (Kii);
Vii) A voluntary act a performed by an agent A is *unintentional* if A does not knowingly perform a;
Viii) Else, a voluntary act a performed by an agent A is *neither intentional nor unintentional*.

And the category of nonvoluntary acts:

NVi) A nonvoluntary act *a* performed by an agent *A* is *intentional* if *a* is an intentional act that *A* performs under duress or out of obligation;
NVii) A nonvoluntary act *a* performed by an agent *A* is *involuntary* if *a* could have been performed voluntarily but was not (Bennett and Hacker, 2003);
NViii) Else, a nonvoluntary act *a* performed by an agent *A* is *neither intentional nor involuntary*.

With the foregoing as background, let us return to the question of what kinds of doings are those of committing suicide and attempting to commit suicide. Now, the concepts that denote these doings can reasonably be seen as technical concepts whose meanings are informed by ordinary language concepts such as *want*, *know*, and *believe*. In my view, this makes them well suited to being described in the style of Kenny's scenarios. To this task I now turn.

11.2.2 Committing suicide

Let: *A* be a human agent; *a* be some act performed by *A* and *a'* its result; *b'* be the result '*A* is dead' and *b* the corresponding act; and $a' \rightarrow b'$ represent the state of affairs that *a'* brings about *b'*.

Now, when *A* performs act *a*, he performs act *b*, the result of which is '*A* is dead.' However, to perform act *b* is not to perform the act of committing suicide unless certain psychological ingredients are present, for, of course, if *a* were, for example, the act of touching a high-voltage wire (or ingesting sleeping pills), the situation, as described, would leave open whether *A*'s death were an accident or a suicide.

Committing suicide is an *intentional* act. Thus, it makes no sense to claim that Agent *A* committed suicide unless he: (a) *knowingly* performed act *a*; (b) *knew* that in performing act *a* he was performing act *b* (foresaw consequence *b'* of act (a); and (c) *wanted* to perform act *b*.

The most bare-bones varieties of the act of committing suicide are those that can be described *in toto* as either a Kenny 1 (in which *A wants* to perform act *a*), or a Kenny 2 (in which he does not want to perform act *a* for any other reason than to perform act *b*) scenario. I will, then, use like terminology and call these types of suicides Scenario 1 and 2 suicides. I note that: (a) on Kenny's analysis, each of the acts *a* and *b* of Scenarios 1 and 2 suicides is an intentional act; and (b) one says that

Agent *A* committed suicide (i.e., performed *b*) by performing *a* (e.g. shooting self, taking an overdose, jumping off of a building).

Although the act of committing suicide is always an intentional act, it can be performed either voluntarily or nonvoluntarily. Moreover, it can be performed either partly or completely for the sake of performing some set of consequent acts. Clearly, then, at least for the purposes of psychological research, there is no sense in portraying the act of committing suicide as unitary. Distinct types of suicides must be distinguished on the basis of Agent *A*s knowledge, beliefs, and wants in respect to his performing of acts *a*, *b*, and any acts related, and consequent, to these.

This is an important point to note when the issue at hand is the formulation of explanatory accounts of *the* act of committing suicide, and, in particular, the employments of certain of the concepts that appear in these accounts. The proper formulation of an explanatory account of the act of committing suicide depends, in part, on the nature of the explanandum, and it turns out that there are manifold types of suicide, hence, manifold potential explananda. I will elucidate two sub-species of the act of committing suicide.

An intentional, nonvoluntary, act is an intentional act that is performed under duress or out of obligation (Bennett and Hacker, 2003). Certain suicides fit this bill. Let: *A* be a human agent; *a* and *b* be acts and *a′* and *b′* = '*A* is dead', their results; $c' = \{c_1', \ldots, c_k'\}$ be a set of consequent states of affairs (results) and $c = \{c_1, \ldots, c_k\}$ the corresponding acts; and $a' \rightarrow b'$. Then, Agent *A* committed a nonvoluntary act of suicide if: (a) he *knowingly* performed act *a*; (b) *knew* that in performing act *a* he was performing act *b* (foresaw consequence *b′* of act *a*); (c) *knew* (in which case $b' \rightarrow c'$) or *believed* (in which case $b' \rightarrow c'$ may or may not have been the case) that in performing act *b* he was performing *c*; (d) *did not want* to perform either of acts *a* or *b* for any other reason than to perform act *c*; and (e) was compelled to perform *c*. I will call *this* type of suicide a Scenario 3 suicide.

The acts {*a*,*b*} of a Scenario 3 suicide are related in the manner described in Kenny's Scenario 2. Provided that $b' \rightarrow c'$ was the case, so too are the acts {*b*,*c*}, the only modification being that, in a Scenario 3 suicide, Agent *A* was *compelled*, rather than merely *wanted*, to perform *c*. If $b' \sim\!\!\rightarrow c'$ was the case, then *A* believed mistakenly that in performing *b* he was performing *c*.

Thus, it may be concluded that the acts *a* and *b* of a Scenario 3 suicide are intentional acts, and act *c* is an intentional act only if it *was*, in fact, performed; equivalently, if, in fact, $b' \rightarrow c'$ was the case (i.e., *A* knew

or believed correctly that, in performing act *b*, he was performing *c*). Example: a failed samurai commits suicide (*b*) by disembowelment (*a*) in order to fulfil the code of seppuku (*c*).

Naturally, an individual's having performed an act of committing suicide will bring into existence a great many states of affairs. Many of which the individual will not have foreseen or wanted. But those that were, if there are any such, are psychologically relevant to an individual's having committed suicide, and, hence, are relevant to a description of such an explanandum. Let: *A* be a human agent; *a* and *b* be acts and *a'* and *b'* = '*A* is dead' their results; $c' = \{c_1', \ldots, c_k'\}$ be a set of consequent states of affairs (results) and $c = \{c_1, \ldots, c_k\}$ the corresponding acts; and $a' \rightarrow b'$. Then, Agent *A* committed a voluntary act of suicide (partly or completely) *for the sake of c'* if: (a) he *knowingly* performed act *a*; (b) *knew* that in performing act *a* he was performing act *b* (foresaw consequence *b'* of act *a*); (c) *knew* (in which case $b' \rightarrow c'$ was the case) or *believed* (in which case $b' \rightarrow c'$ either was or was not the case) that in performing act *b* he was performing act *c*; and (d) *wanted* to perform act *c*.

If acts *b* and *c* are related in the manner described in Kenny's Scenario 1, the suicide was *partly* for the sake of *c'*. If they are related in a manner described in Kenny's Scenario 2, the suicide was *completely* for the sake of *c'*. I will call these types of suicide, Scenario 4 and 5 suicides, respectively. Once again, on Kenny's analysis, acts *a* and *b* are intentional acts, and act *c* is an intentional act if it was, in fact, the case that $b' \rightarrow c'$ (i.e., *A* knew or believed correctly that, in performing act *b*, he was performing act *c*). Examples: a man kills himself (*b*) by taking poison (*a*) because he is sick of living, wants to die, and believes that, in killing himself, he will punish the object of his unrequited love (*c*), and wants to do so (Scenario 4); a man whose criminal ways have been detected shoots himself (*a*), fatally (*b*), in order to avoid suffering shame (c_1) and being prosecuted (c_2) (Scenario 5).

11.2.3 Attempting to commit suicide

What kind of doing is that of *attempting suicide*? When an Agent *A* has decided to perform an act of committing suicide, perhaps bringing to a close a period of indecision, he has formed an intention to commit suicide (cf. Bennett and Hacker, 2003). He *knows* what it is to perform this act (foresees the result of the act he intends to perform) and *wants* to bring about the result '*A* is dead.' Of course, the fact that *A* has formed an intention to perform the act of committing suicide does mean that he will do so. He must not only take initiative, but actually *succeed* in performing the act.

As with the sentence '*A* will, next Sunday, attempt to break four minutes for the mile', the sentence 'Agent *A* will attempt to commit suicide tomorrow at 5:00 p.m.' heralds the performance of an act while acknowledging the possibility that *A* might not succeed in its performance. On the other hand, the sentence 'Agent *A* attempted suicide' indicates a *failed* performance of the act of committing suicide, or, in other words, the failure of *A* to bring about a particular state of affairs.

Let: *A* be a human agent; *a* be some act and *a'* its result; *b'* be the result '*A* is dead' and *b* the corresponding act; and $a' \sim\!\!\rightarrow b'$ represent the state of affairs that *a'* does not bring about *b'*. Then Agent *A* attempted suicide if: (a) he *knowingly* performed act *a*; (b) *believed* (mistakenly) that in performing act *a* he was performing act *b*; (c) *wanted* to perform act *b*.

We say that Agent *A* attempted suicide by performing act *a*, or, equivalently, that act *a* was an attempted suicide. Now, save for Agent *A*'s mistaken belief that in performing act *a* he was performing act *b*, the logical structure of attempted suicide is identical to that of suicide itself. Thus, the act *a* of attempting to commit suicide can be performed: (a) voluntarily or not; (b) partly or completely for the sake of performing some set of acts that Agent *A* knew or believed he would be performing in performing the act of committing suicide. Common to the manifold types of attempted suicides are the facts that: (a) act *a*, the attempted suicide, is an intentional act and (b) neither act *b*, nor any acts that Agent *A* knew or believed that he would be performing in a successful performance of act *b*, were, in fact, performed.

11.3 The problematic concepts

Let us now return to the focal problem of the chapter, the conceptual confusion that attends the employments of the concepts *intention*, *motive*, and *reason* in the context of the formulation of explanatory accounts of the phenomena of suicidology. Two of these phenomena, explananda of the explanatory accounts formulated by suicidologists, are the acts of committing suicide and attempting to commit suicide. As we have seen, each of these explananda is actually a manifold class of acts.

In the case of committing suicide [attempting to commit suicide], the act *b* [*a*] (the act of committing suicide) [the act of attempting to commit suicide] that must be explained is, for scenarios of types 1, 2, 4, and 5, a voluntary, intentional act, and, for a scenario of type 3, a nonvoluntary, intentional act. Thus, the task will be to clarify the correct employments

of the problematic concepts when they are ingredients of explanations of why an Agent *A* performed either a particular voluntary, intentional act or a nonvoluntary, intentional act.

The grammars of psychological concepts are ramifying, and the meanings of the concepts *reason*, *intention*, and *motive* are linked in complex ways to the meanings of other concepts. A clarification of the concept *reason* begs for a contrasting of it with the concept *cause*, a clarification of *motive* requires clarity in respect to the meanings of the concepts *emotion*, *appetite*, and *agitation*, and the concepts *intention* and *intentional act* (as we have, in the latter case, already seen) are related in complex ways to the concepts *aim*, *want*, and *knowledge*. I will find it most convenient to order my clarifications of the problematic concepts in the following way: (a) reason (and cause); (b) motive (and emotion, appetite, agitation, and other supporting concepts); and (c) intention (and intentional act).

11.3.1 Reason

Reasons are potential ingredients of explanations of *certain* types of *acts*. More particularly, an act is a doing that an Agent *A* performs or fails to perform, and, for certain types of doing that an Agent *A* performs or fails to perform, reasons are potential ingredients of answers to questions of *why A* did ξ. Within the category of acts, it is paradigmatically the voluntary act that is open to explanation in terms of reasons.[1] This is because *A*'s performing of a voluntary act γ involves *his* exercising of a two-way power to do or refrain from doing γ (Bennett and Hacker, 2003), and the rules of language establish as a possibility that *A* had reasons for having done, or refrained from having done, γ. In the event that *A* did γ, rather than having refrained from doing so. In contrast, *A* cannot have had reasons for having slipped and fallen, for slipping and falling are not acts, let alone acts the performance of which involved *A*'s exercising of a power to do or refrain from doing.

Let it be the case that an Agent *A* has performed a voluntary act γ. Then:

Ri) it is a possibility (allowable under the rules of language) that *A* had reasons for having performed γ;
Rii) there is no necessity that *A* had reasons for having performed γ.

If *A* did, in fact, have reasons for having performed γ, then:

Riii) these reasons are (non-causal) factors that *A* considered to be relevant to his having performed γ;

Riv) these reasons are *his* reasons. The possessive grammar, here, marks an '…asymmetry between first- and third-person, present tense, psychological statements' (cf. Baker and Hacker, 1982). *Clarification*: The reasons that *A* provides for having performed γ are standardly groundless (not resting on any *evidence*) avowals of his wants, desires, beliefs, and the like. *A* may choose to avow his reasons or he may choose to keep them to himself. He may mull over his reasons, modify them or change them outright, or come to have a new set of reasons for having performed the same act.

On the other hand, a third-person ascription to *A* of reasons for *his* having performed γ is made on the basis of behavioural criteria; that is, behaviours of *A*, including *A*'s avowals, that constitute grammatical justification for the claim that *A* had these reasons for having performed γ (cf. Baker and Hacker, 1982).

While both *A*'s truthful first-person avowals and third-person ascriptions of predicates to *A* on the basis of behavioural criteria confer grammatical certainty upon the relevant judgments made about *A*, grammatical certainty is not the same thing as absolute, unconditional, certainty, and judgments made on the basis of either avowals or correct third-person ascriptions are logically defeasible (under, for example, a broadening of the circumstances or a demonstration of *A*'s pretence) (Baker and Hacker, 1982);

Rv) The richness of the explanations of γ *possible* through a citation of *A*'s reasons for having performed γ will standardly vary as a function of the *type* of voluntary act that γ is. *Clarification*: In the case of voluntary acts that are either unintentional, or neither unintentional, nor intentional, the prospects will be limited. Why was *A* running his fingers through his hair while he studied algebra? That is, why was he performing *that* voluntary, unintentional act? Answer: He says that he wasn't aware that he was doing so. Why did *A* crush blades of grass on his way to the car? That is, why did he perform that particular voluntary act that was neither intentional nor unintentional? Answer: He had not even considered the point, but, in any case, had to do so in order to get to the car.

The logical structure of voluntary, intentional acts, on the other hand, opens the way for richer explanations involving *A*'s reasons for having done. If *A* has performed a voluntary, intentional act φ, *A knowingly*

performed φ (he foresaw result φ′) and performed φ either solely because he *wanted* to bring about result φ′ or because he *knew* or *believed that*, in performing φ, he was performing some consequent act whose result he *wanted* to bring about. Act φ is, by virtue of its being a voluntary, intentional act, tied to factors, namely, *A*'s wants, beliefs, and desires, that were relevant to *A* in his having performed φ □

11.3.1.1 Reason and cause

The concepts *reason* and *cause* are not synonymous. Although both causes and reasons are potential ingredients of explanations of the doings of agents, they are ingredients of explanations of *different* types of doings. As we have noted, *A*'s reasons for having performed a voluntary act γ, should he have had reasons for having done so, are legitimate ingredients of explanations of γ. Causes are not; voluntary acts performed by agents *have no causes*.

If the doing in question is, on the other hand, a nonvoluntary act ρ, ρ may have a cause ρ_c, but the agent *A* who performed ρ cannot have had reasons for having done so. This is because in having performed ρ, *A* was not exercising a two-way power to do or refrain from doing ρ.

If a nonvoluntary act ρ has a cause ρ_c:

Ci) ρ_c is the thing that brought about ρ;
Cii) the occurrence of ρ can be explained with reference to ρ_c; hence, ρ_c is a legitimate ingredient of explanations of ρ.

Though it would make no sense to inquire as to the *reasons A* had for slipping and falling, it would make perfect sense to investigate the conditions under which the accident occurred, the aim being to discover its *causes*. And though it would, similarly, be nonsensical to inquire as to *A*'s *reasons* for the involuntary twitching of his face (the twitching of *A*'s face not under control of a two-way power possessed by *A* to do or refrain from doing), it would make perfect sense for *A* to seek out medical assistance in the hope of coming to know the cause of his twitch.

One inquires as to *why* Agent *A* performed voluntary act γ (and, if *A* had reasons for having done, and these reasons are known, cites these reasons), but *what* is the cause of ρ, say, his involuntary twitch. Whereas *A*'s reasons for performing the voluntary acts he performs are *his* reasons (are formulated by him out of his knowledge, beliefs, and wants), the cause ρ_c of ρ, is not *A*'s cause. *A* has reasons, not causes. *A* may *avow* his reasons for performing a voluntary act, but not the cause of his twitch. Causes are as they are. They are independent of what *A* knows,

believes, and wants. It is incoherent to say that *A* has reasons for having performed an act but does not know what these reasons are, but not at all to say, for example, that *A*'s twitch has a cause, but *A* does not know what it is.

11.3.2 Motive

Motives are potential ingredients of explanations of intentional acts (Bennett and Hacker, 2003). Let it be the case that an Agent *A* has performed an intentional act β. Then:

Mi) it is a possibility (allowable under the rules of language) that, in performing β, *A* was *acting out of* a motive ζ (Bennett and Hacker, 2003);

Mii) there is no necessity that, in performing β, *A* was acting out of a motive.

If, in performing β, *A* was acting out of a motive ζ, then:

Miii) ζ is an appetite, emotion, or desire of *A*'s; more generally, ζ is a psychological phenomenon whose denoting concept has a formal or formal and specific object. *Clarification*: Appetites such as *hunger*, *thirst*, and *lust* are blends of sensation and desire (Bennett and Hacker, 2003). Sensations do not have objects, but desires do. The desire component of an appetite has a formal, but not a specific, object. The formal object of *hunger* is food/nutritional sustenance, of *thirst* is liquid refreshment, and of *lust* is sexual intercourse (Bennett and Hacker, 2003).

Emotions such as *grief*, *love*, *resentment*, *fear*, and *jealousy* have both formal and specific objects (Bennett and Hacker, 2003). The formal object of one's jealousy is the thing that makes one jealous, the specific object, some particular state of affairs; example, that the object of one's affections was seen entering a house with a former flame.

Finally, a desire is identified by the object κ that satisfies it. The sentence 'I desire a steak for dinner' specifies the the object of my gustatory desire is steak.

Although conceptually closely related to the appetites, emotions, and desires, the agitations and moods do *not* have objects, and so cannot play the role of motive for an intentional act (Bennett and Hacker, 2003). Agitations such as being *thrilled*, *excited*, *shocked*, *terrified*, or *horrified*: A(a) are short-term affective disturbances that temporarily *inhibit* motivated

action. One may behave in a certain way *because* one is terrified, but not *out of* terror; A(b) are modes of reaction. One recoils *with* revulsion, shrieks *in* terror, cries out *in* horror (Bennett and Hacker, 2003). Moods such as feeling *cheerful*, *depressed*, *contented*, or *euphoric*: MO(a) are frames of mind that are either occurrent states or longer-term dispositional states; MO(b) colour one's thoughts and pervade one's reflections,[2] and MO(c) are exhibited not in patterns of action but, rather, in the manner in which one does whatever one does (e.g. in one's demeanour or tone of voice) (Bennett and Hacker, 2003).

Miv) the motive (appetite, emotion, or desire) ζ out of which *A* was acting in his performance of β was the (non-causal) origin of a *pattern of behaviour* (sequence of voluntary, intentional acts of which β was an element) oriented towards the object of ζ (Bennett and Hacker, 2003). *Clarification*: Appetites, emotions, and desires can be motives for intentional acts because they have (formal or formal *and* specific) *objects*, and it is characteristically human that individuals deal with the objects of emotions, desires, and appetites through the performance of sequences of intentional acts.

When, for example, an Agent *A* is in love, the object of his love is some individual, say, individual *B*. *A* must deal, in some way, with his passively acquired feelings of love for *B*, and may well do so by performing intentional acts: example, by acting in a manner that he believes will please *B*, seeking out *B*'s company, spending time with *B*, attempting to woo *B*, buying gifts for *B*, or writing poetry for *B*. His love of *B* is the motivating emotion of a sequence of intentional acts, a pattern of behaviour. In performing the intentional acts which this pattern comprises, *A* is *acting out of his love for B*;

Mv) With respect to the issue of asymmetry between first- and third-person, present tense, utterances, the concept *motive* is a mixed case (cf. Baker and Hacker, 1982, p. 237). The ascription to *A* of motive ζ for his having performed β rests on a demonstration that β is an element of a sequence of intentional acts performed by *A*, these acts oriented towards the object of ζ. Although there is no first- and third-person asymmetry attendant to utterances about the acts themselves (*A* does not *avow* performance of an act), there *is* such an asymmetry attendant to the ascription of appetites, emotions, and desires (i.e., *A* may avow his appetites, emotions, and desires).

11.3.3 Intention

Agents *form* intentions to perform voluntary, and nonvoluntary, intentional acts φ, *manifest* these intentions in nonlinguistic behaviour, and express them in language. Let φ be a voluntary or nonvoluntary intentional act performed by an agent *A*. Then:

Ii) It is a possibility (allowable under the rules of language) that *A* formed, prior to having performed φ, an intention to perform φ;
Iii) There is no necessity that *A* formed an intention to φ prior to having performed φ.

If *A* did, in fact, form an intention to perform φ prior to his having performed φ, then:

Iiii) His intention was neither a state of mind, nor an inner phenomenon, nor a private experience, nor a *thing*. *A manifests* his intentions in his nonlinguistic behaviour and expresses them in language. When *A* states, 'I am going to φ', he expresses his intention to φ by heralding the performance of φ. When *A* declares, 'I have decided to φ', he indicates that a period of indecision has been brought to a close and an intention to φ, formed. When *A* employs the expression 'I intend to φ', he is *avowing* an intention to φ;
Iiv) *A*'s manifesting of his intention to φ in his nonlinguistic behaviour, or expressing of it in language, gives an addressee '...a (non-inductive) ground for prediction' (Bennett and Hacker, 2003, p.103) or, in other words, a grammatical basis for expecting that *A* will perform φ;
Iv) The relation between his intention to φ and φ is grammatical rather than empirical. In particular: (a) his performing of φ *satisfied* his intention to φ; (b) his intention to φ was not a cause of (did not bring about) φ; (c) *Nothing* in *A*'s having formed an intention to φ necessitated that *A* satisfy this intention by performing φ. Having formed an intention to φ, *A* still had to, and did, in fact, both take initiative and successfully perform φ, in order that his intention to φ was satisfied.

As with the concept *reason*, the concept *intention* is marked by a first-person/third-person asymmetry. When an Agent *A forms* an intention to

φ, he may either avow this intention or keep it to himself. Third-person ascriptions of intentions to *A* rest on behavioural criteria, that is, *A*'s 'behaviour in context' (Baker and Hacker, 1982, p.235).

11.4 The explanations of suicidology

The phenomena of suicide and attempted suicide, whose explanations are at issue, are manifold classes of intentional acts, either voluntarily or nonvoluntarily, performed by agents. Let φ* stand for an intentional act belonging to one of these classes, and let it be the case that an Agent *A* has performed φ*. It may, then, be concluded that:

i) it is a possibility (allowable under the rules of language) that *A* had reasons for having performed φ*. If *A* did, in fact, have reasons for having performed φ*, these reasons are legitimate ingredients of explanations of why *A* performed φ*. It follows, then, that the author of any such an explanation must be able to ascribe to *A*, *A*'s reasons for having performed φ*, and this will require that he is either privy to *A*'s avowals of his reasons (say, as a consequence of his having been present during these avowals, or in possession of letters, diary materials, or other writings, in which they were expressed), or other criterial evidence that supports their ascription to *A* in the third-person mode;

ii) it is a possibility (allowable under the rules of language) that, in performing φ*, *A* was acting out of a motive ζ. If, in performing φ*, *A* was acting out of a motive ζ, then motive ζ is a legitimate ingredient of explanations of why *A* performed φ*. The sense in which ζ *explains* φ* is that φ* is an element of a sequence of intentional acts that originated in ζ (an appetite, emotion, or desire), the acts that constitute the sequence oriented towards the object of ζ. Hence, an explanation of φ* in terms of ζ will take the form of a description of both the originating appetite, emotion, or desire, and the subsequent sequence of intentional acts that contains φ*;

iii) it is a possibility (allowable under the rules of language) that *A* formed an intention to φ* prior to his having performed φ*. If, in fact, *A* formed an intention to perform φ* prior to his having performed φ*, this fact is not *relevant* to an explanation of *A*'s having performed φ*, the reason being that the fact that *A* intended to φ* prior to having done so does not bear on the question of *why* he did so. Knowledge of *A*'s intentions *is*, however, relevant to the formulation of *predictions* of *A*'s future behaviour;

iv) because φ* is not caused, and so does not *have* causes, a citation of hypothesized causes is *not* a legitimate constituent of explanations of why *A* performed φ*.[3]

11.4.1 An example

The popular media have documented many cases of individuals whose suicides have brought to a close a period during which they had been suffering through the agonies of unrequited love for another.[4] In one such recent case,[5] a fourteen-year-old boy from Abergele, Wales (hereafter, *A*), met a fifteen-year-old girl from Huddersfield, England (hereafter, *B*) at Disneyland Paris and fell in love with her. While waiting in line for a ride, *A* and *B* chatted and exchanged personal information. Following their returns to their respective homes, *A* and *B* corresponded by e-mail and text message. In these exchanges, *A* described his love for *B* as follows: 'Words don't seem able to come out of my mouth when it comes to you ... after I heard your voice everything seemed to flood out ... When we spoke I knew you were the girl I wanted to be with and my heart fluttered, but after you left a chill fell into the gap and I felt so alone, so stranded, so longing to be with you ... All I can think about is how amazing you are.' *A* also alluded to problems in his life, notably, a home situation in which there had recently been much arguing. *A*'s attempts to arrange to meet with *B* in Manchester came to nothing, because *B* believed that her parents would forbid such a meeting.

A committed suicide by hanging himself, leaving behind notes for *B*, his family, and friends. In the note to *B*, he stated, 'I always said I would give my life just to see you again, but now I'm giving my life not to see you again', 'I know you will be upset ... but I want you to know I will be in heaven watching over you always.'

We have then:

i) an originating emotion, *E*, *A*'s feelings of love for *B* that: (a) is a composite of an emotional attitude that informed *A*'s life from the time of his having met *B* until he ended his life by committing suicide and episodic emotional perturbations[6] (see Bennett and Hacker, 2003, pp. 203–7); (b) has a specific object, that, of course, being *B*.

I will symbolize the originating emotion, the object of which was *B*, as (*A*)*E*(*B*);

ii) a sequence of intentional acts, d_t, $t = 1..k$, performed, subsequent to the onset of $(A)E(B)$, by A, and oriented towards B. These acts included A's sending to B e-mails and text messages, attempting to convince her to meet with him in Manchester, and writing her a suicide note.

Let this sequence be symbolized as $\{d_1(B), d_2(B), \ldots d_k(B)\}$;

iii) an act b of committing suicide performed by A subsequent, of course, to $\{d_1(B), d_2(B), \ldots, d_k(B)\}$. Act b was: (a) a Scenario 4 or 5 suicide, for it is clear that it was performed by A at least in part for the sake of bringing about certain consequent states of affairs, $c' = \{c'_1, \ldots, c'_p\}$, having to do with B. In particular, A seems to have believed that his committing suicide would bring about a state of eternal closeness with B; (b) performed by A through his performing of the intentional act a of hanging himself.

Ordering these elements temporally, we have:

$$\begin{array}{rccc} & a' \rightarrow & b' \rightarrow & c' \\ & \uparrow & \uparrow & \uparrow \\ (A)E(B)\ldots\{d_1(B), d_2(B), \ldots, d_k(B)\}\ldots & a & b & c \end{array}$$

Now, in this case a motive is a legitimate constituent of an explanation of why A performed the act b of committing suicide: A was *acting out of* his love for B. What warrants ascription of such a motive to A, that is, justifies the claim that, in performing the act of committing suicide, A was acting out of his love for B, is the existence of a *pattern of behaviour* that originated in E, terminated in the explanandum, b, and was oriented towards B.

The assistant deputy corner for North Wales concluded that 'The only clue that there is to what might have been going on in [the boy's] mind comes from the e-mails and text messages and the note left by him for this girl he met in Disneyland ... *there is nothing at all to give a reason why* [italics added].' This is mistaken. As I have noted, aside from the fact that A was acting out of a motivating emotion, A avowed at least one reason for his having committed suicide, and this reason is a legitimate part of an explanation of why he committed suicide. In his note to B, he revealed that he viewed it as *relevant* to his intention to commit suicide, his belief that, in so doing, he would be bringing about a state of eternal closeness with her.

Various authorities portrayed both *A*'s difficulties at home and his having been called 'gay' by peers earlier in his life as possible *causes* of his having committed suicide. This is mistaken. The act of committing suicide that *A* performed was an *intentional act*, and so *had* no causes. Of course, *A* may have considered these to be *reasons* for his having committed suicide; he did, after all, take the trouble of mentioning to *B*, the object of his love, his familial difficulties.

Finally, the fact that *A* wrote suicide notes is criterial support for the claim that, prior to committing suicide, he had formed an intention to do so. However, that an intention to commit suicide can be rightly ascribed to *A* does not have any relevance to the issue of *why* he committed suicide.

Notes

1. But not exclusively so; nonvoluntary, intentional acts are also open to being explained with respect to A's reasons for having done.
2. For example, a depressed individual, i.e., one who is suffering from depression, is suffering from a dispositional depression or a long-term proneness to feeling depressed (a proneness to having his thoughts and reflections be coloured black). Whether or not there is a biological basis to his proneness to feeling depressed is an empirical issue that is currently unresolved.
3. Once again, certain constituent elements of a scenario in which A performs φ* (e.g. bodily states and afflictions, contractions of muscles involved in A's performing of φ*) may have causes, which would need to be cited in any explanations that are formulated of these elements.
4. Interestingly, the protagonist of Goethe's *The Sorrows of Young Werther*, published in 1774, commits suicide after suffering unrequited love for a married woman.
5. The details of this case I know only as reported in a story by Liz Hull, *The Mail*. As I am not attempting to provide an accurate characterization of this particular case, but only a context in which to demonstrate certain of the conceptual clarifications made in the current manuscript, any distortions I visit upon the case as a result of my uncertain understandings of its details should be of no consequence.
6. It is hard to imagine love in the absence of such perturbations, but if the reader is looking for criterial evidence for this ascription, note that A refers, in his notes to B, to the flutterings of his heart, a chill descending over him in B's absence, and, more generally, feelings of longing and loneliness.

References

M. Alicke, M. Weigold, and S. Rogers (1990) 'Inferring Intentions and Responsibility from Motives and Outcomes: Evidential and Extra-Evidential Judgments', *Social Cognition*, 8(3), 236–305.

B. Baker, G. and P. M. S. Hacker (1982) 'The Grammar of Psychology: Wittgenstein's *Bemerkungen Über die Philosophie der Psychologie'*, *Language and Communication*, 2(3), 227–44.

H. Bancroft, A. Skrimshire, and S. Simkin (1976) 'The Reasons People Give for Taking Overdoses', *British Journal of Psychiatry*, 128, 538–48.

A. Beck, D. Schuyler, and I. Herman (1974) 'Development of Suicidal Intent Scales' in A. Beck, H. Resnick, and D. Letieri (eds) *The Prediction of Suicide* (Philadelphia: The Charles Press).

M. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Malden, Mass.: Blackwell Publishing).

J. Birtchnell and J. Alarcon (1971) 'The Motivation and Emotional State of 91 Cases of Attempted Suicide', *British Journal of Medical Psychology*, 44, 45–52.

J. Hettiarachchi and G. Kodituwakku (1989) 'Self Poisoning in Sri Lanka: Motivational Aspects', *The International Journal of Social Psychology*, 35(2), 204–8.

H. Hjelmeland and B. Knizek (1999) 'Conceptual Confusion about Intentions and Motives of Nonfatal Suicidal Behaviour: A Discussion of Terms in the Literature of Suicidology', *Archives of Suicide Research*, 5, 275–81.

L. Hull (2008, October 2) Heartbroken boy, 14, hanged himself after unrequited girl he met on school trip. Mail Online. Retrieved December 12, 2012, from http://www.dailymail.co.uk/news/article-1065954/Heartbroken-boy-14-hanged-unrequited-love-girl-met-school-trip.html

A. Kenny (1966) 'Intention and Purpose', *The Journal of Philosophy*, 63(2), 642–51.

M. Kovacs, A. Beck, and A. Weissman (1975) 'The Use of Suicidal Motives in the Psychotherapy of Attempted Suicides', *American Journal of Psychotherapy*, 29(3), 363–8.

N. Lukianowicz (1972) 'Suicidal Behaviour: An Attempt to Modify the Environment', *British Journal of Psychiatry*, 121, 387–90.

R. Maris, A. Berman, and M. Silverman (2000) 'The Theoretical Component of Suicide' in R. Maris, A. Berman, and M. Silverman (eds) *Comprehensive Textbook of Suicidology* (New York: Guilford), 26–61.

A. Reber and E. Reber (1985) *Penguin Dictionary of Psychology* (London: Penguin Books Ltd).

J. Schulte (1993) *Experience and Expression: Wittgenstein's Philosophy of Psychology* (London: Oxford University Press).

M. ter Hark (1990) *Beyond the Inner and the Outer: Wittgenstein's Philosophy of Psychology* (London: Kluwer Academic Publishers).

C. Trevarthen (1982) 'The Primary Motives for Cooperative Understanding' in G. Butterworth and P. Light (eds) *Social Cognition: Studies of the Development of Understanding* (Brighton: Harvester Press), 77–108.

V. Velamoor and Z. Cernovsky (1992) 'Suicide with the Motive "To Die" or "Not To Die" and Its Socioanamnestic Correlates', *Social Behaviour and Personality*, 20(3), 193–8.

G. H. Von Wright (1963) *Norm and Action: A Logical Enquiry* (New York: Humanities).

12

The Neuroscientific Case for a Representative Theory of Perception

John Preston and Severin Schroeder

It has been urged repeatedly over the last two decades that empirical findings in neuroscience and psychology provide compelling reasons for endorsing a representative theory of perception. Richard L. Gregory and John R. Smythies are perhaps the best-known advocates of this view. When it comes to vision, in particular, scientists of this persuasion think that the supposed alternative, 'direct realism', is hopelessly naïve, and they conclude that, as Francis Crick puts it, 'What you see is not what is *really* there; it is what your brain *believes* is there' (Crick, 1994, p. 31). We will take a critical look at some of these empirical findings, and discuss the extent to which they support the more sweeping philosophical claims scientists have drawn from them, in particular the advocacy of representative theories of perception.

In the decades of the mid-twentieth century, leading figures from the philosophical tradition of conceptual analysis, such as Ludwig Wittgenstein, J. L. Austin, Gilbert Ryle, and Alan R. White, began their pioneering project of analysing our extensive panoply of perceptual concepts. In more recent years, philosophers such as Peter Hacker and John Hyman, strongly influenced by Wittgenstein's work, have continued this project, deepening and sophisticating such analyses, and drawing attention to flaws in causal and representative theories, whether those theories originate from scientists or from philosophers.[1] The extent to which the scientific advocates of the representative theory have failed to learn the lessons of this work (even in the very formulation of their opponents' views) will soon emerge.

12.1 Smythies and Ramachandran

The esteemed neuroscientist John R. Smythies is a long-time advocate of the representative theory of perception. Since his first book, *Analysis of Perception* (Smythies, 1956), he has produced a series of publications (Smythies, 1965, 1993, 1994, 2005, 2009) in which he has tirelessly searched for arguments, many of them drawn from experimental considerations, to refute what he takes to be the alternative view, which he calls 'direct realism'. In 1997, he teamed up with another greatly-esteemed neuroscientist, Vilayanur S. Ramachandran, to write a short paper entitled 'An Empirical Refutation of the Direct Realist Theory of Perception'. It is this work on which we shall focus here, for it manages to raise many of the key issues with which we are concerned.

12.2 The characterization of 'direct realism'

Smythies and Ramachandran (1997) argue that a particular set of recent experiments refute 'direct realism' and support the rival 'scientific Representative Theory'. The way they set the debate up bears some consideration. They begin by characterising 'direct realism' as the view that 'the visual field contains the physical object itself, and thus the phenomenal object is identical to the physical object' (Smythies and Ramachandran, 1997, p. 437). There are several things of note here.

One of them is the casual reference to 'physical objects', as if everything one could be said to see falls into this category. Our 'direct realist' has not learnt one of the main lessons of J. L. Austin's (1962) *Sense and Sensibilia,* where the use of such a category as a single catch-all term for whatever is thought to be perceived is shown to be a liability. Austin showed how advocates of representative theories of perception introduce such categories (in those days it was 'material objects') as stooges, later to be supplanted by their preferred 'mental representations' (some of which are what in Austin's day were called 'sense-data'). Our 'direct realists' would indeed be naïve to frame their view in such terms. The most they should concede here is that in vision, many of the things we perceive count as physical objects.

A second thing to note is the references to the 'visual field' and to 'phenomenal objects'. What can these expressions mean here? Presumably, since they occur in the formulation of 'direct realism', they must refer to features that the 'direct realist' believes in. They cannot, therefore, have built into them the things that such 'realists' regard as

problematic, or at least as not playing an essential intermediary role in vision, such as 'internal' or 'mental' representations.[2] Let us take it, then, that 'the visual field' refers to all that comes into view when the eyes are turned in some direction, and that 'the phenomenal object' refers to the sorts of things people *think* they see, or report themselves as seeing. If we combine this with the above point about the expression 'physical objects', the resulting view is thus that physical objects feature among the things one can see, as well as among the sorts of things people think they see, or report themselves as seeing.

Thus characterised, though, 'direct realism' is entirely commonsensical, and not a theory at all. Smythies and Ramachandran's (1997, p. 437) opening remark, to the effect that 'Most philosophers today support the Direct Realist theory of perception', nevertheless surely contains *some* truth, on this reading, but this would hardly mean that 'most philosophers' keep bad company in this respect, for it really would be surprising if Smythies and Ramachandran had come across an experiment that refutes *this* view.

Common sense, though, does not contain the particular contrast between 'direct' and 'indirect' vision that representative theorists require. Instead, that contrast gets its sense from the opposition between direct realism and representative theories of perception, according to which physical objects (and other public phenomena, of course) are either *not* themselves seen, or are seen only in virtue of our seeing or experiencing phenomena of an entirely different kind, *mental* phenomena (such as sense-data, or 'mental representations').

12.3 The Kovács experiment

Smythies and Ramachandran (1997) go on to describe the experiment that they think refutes 'direct realism', which was performed by Ilona Kovács and her collaborators:

> Their procedure was to take two pictures – A showing a group of chimpanzees and B showing the dense foliage of a tropical jungle. Each subject-matter completely filled the picture. They then constructed two further pictures, C and D. Each of these was made up of a patchwork of portions of A interspersed with portions of B with exact fitment. If one eye of a subject is now exposed to picture A and the other eye to picture B, then the subject will of course see first A, then B, then A, then B and so on in typical retinal rivalry. (Kovács, Papathomas, Yang and Fehér, 1996, pp. 437–8)

The pictures in question, A, B, C and D are as follows:

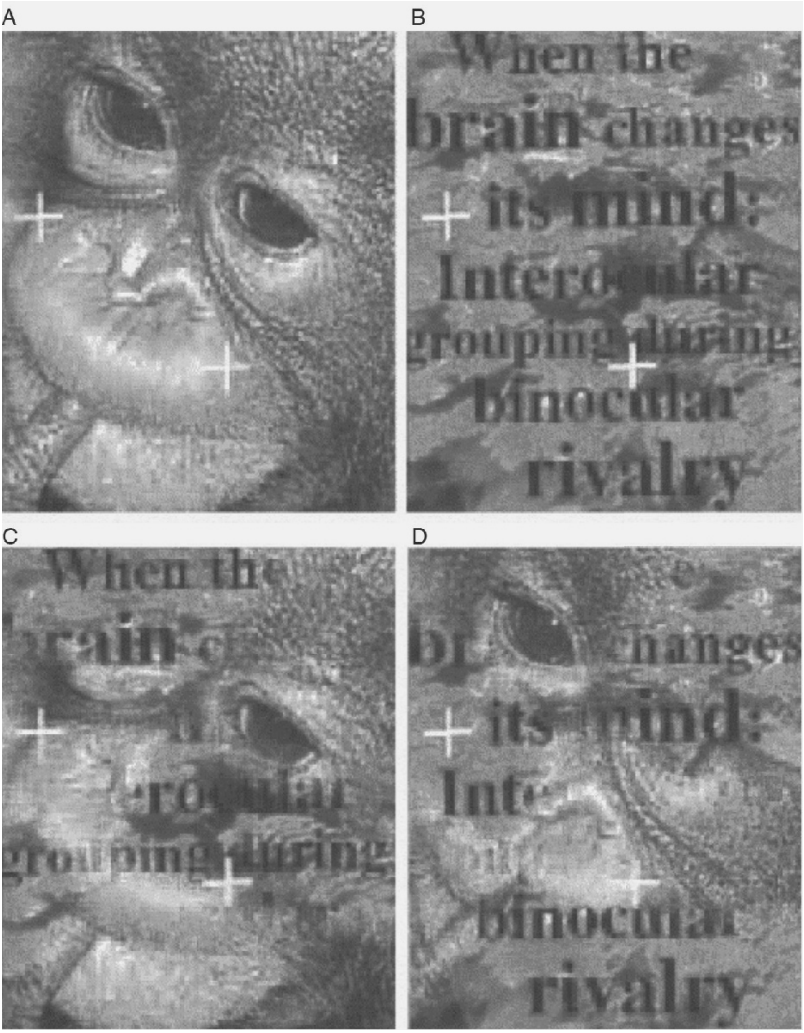


Figure 12.1 Dichoptic pairs, reprinted with permission of Ilona Kovács (Copyright 1996 National Academy of Sciences, USA)

Smythies and Ramachandran (1997) then ask: What will the theory predict if we expose one of the subject's eyes to picture C and the other to picture D? According to them, direct realism must predict that the subject would see the patchwork pictures, C and D, in retinal rivalry since, after all 'that is ... what is out there' (p. 438). In fact, however, the result of the experiment is that subjects report seeing the two original and complete patterns, the monkey and the jungle, A and B, alternating with one another (this being what is referred to as *binocular rivalry*).

It is notable that from their experiment, Kovács, Papathomas, Yang, and Fehér (1996) themselves draw no general conclusion about perception. Their conclusion is that 'binocular rivalry can be driven by pattern coherency, not only by eye of origin' (p. 15511). In other words, the extent to which two natural and coherent patterns are embodied in a patchwork way is a strong factor in determining whether experimental subjects will report seeing the original natural and coherent patterns alternating with one another. For those without a philosophical axe to grind, the conclusion to be drawn is that subjects whose eyes are each presented with one of two different patchwork patterns that fit one another report that they are seeing the corresponding alternating natural and complete patterns. Smythies and Ramachandran (1997), though, take the experiment to show 'that we do not really see what actually is out there', but only representations constructed by the brain.

12.4 The return of the Argument from Illusion

Kovács and colleagues' (1996) experiment has two crucial aspects that both play a role in Smythies and Ramachandran's (1997) argument:

i) The experimental set up produces an illusion.
ii) The illusion is due to a mechanism that facilitates detecting familiar shapes or patterns, but malfunctions under very anomalous conditions.

Smythies and Ramachandran (1997) take (i) already to refute direct realism: 'The Direct Realist Theory would have to predict that the subject would see ... what is out there. However, that is not what happens' (1997, p. 438). In other words, they take direct realism to be committed to the extreme view that perception is infallible: that we always see only what is actually there. This makes the position they attack a straw man, quite different from the direct realism of common sense. When,

commonsensically, we take ourselves to see trees and houses (and not just their mental representations), we do not, for that matter, deny that sometimes we misperceive or even hallucinate.

Smythies and Ramachandran's (1997) line of argument is not new. It is, in fact, that old chestnut, the Argument from Illusion, employed from the 1930s onwards by H. H. Price, A. J. Ayer and others to establish the existence of sense-data. Paradigmatically, the argument runs thus:

(1) Sometimes we're deceived by delusions.
    In such cases we don't see the object we take ourselves to be seeing.
(2) Yet we do see something.
∴ (3) We see sense-data.
(4) Veridical perceptions are indistinguishable from delusions (or else we wouldn't be deceived). (The subjective experience is the same.)
∴ (5) In veridical perception too, we see only sense-data.

The flaws in this argument have long been pointed out (by Austin and by Anthony Quinton, for example).[3] Basically, it rests on an equivocation between two concepts of seeing, and the naïve assumption that what happens in the normal case must be essentially the same as what happens in anomalous cases. The concept of seeing that is most familiar from ordinary contexts is a *factive* concept, for its use carries the implication that what is seen by the subject must actually be present in that subject's environment. Austin and Ryle were particularly keen to stress this feature of our ordinary perceptual verbs, their being *success*-verbs. Seriously to report that someone (oneself or someone else) sees something, Φ, carries the implication that if Φ was not there to be seen, one's report needs revising. The revision in question, of course, is to the effect that the person concerned *thought* they saw Φ. But alongside this, and also present in some ordinary contexts, are what we might call *subjective* uses of seeing, uses in which to say what a subject (oneself or another person) sees is to give expression to what shapes and colours they are visually aware of, with no implication that these are the shapes and colours of things that are present in the subject's external environment.

From a neurophysiological point of view it may be true that what happens in these two kinds of cases is very similar: perhaps, during an hallucination or a vivid dream, the same parts of the brain are activated

that are activated in visual perception. That is irrelevant, though, since those identical neurophysiological processes are not what we see. To think otherwise – to think that the real seeing, including the object really seen, happens inside the brain – is an instance of the homunculus fallacy.[4]

But cannot one's subjective visual experience be the same in both cases? Yes, but that does not mean that what is *seen* is the same in both cases. Under the ordinary concept of seeing, what is seen can also be overlooked, can also exist unseen, or can be misperceived, and so the object of sense perception is independent of our perceiving it. Representative theorists take their cue from the other, more marginal, subjective uses of the verb to see, in order to develop a very different concept of seeing, to which they then switch (in premise (2) above, for example). With this concept, which we might call 'internal-seeing' (and whose object we might call 'one's visual experience'), *esse* is *percipi*: the object seen exists only when it is seen, and has no qualities apart from those that are seen.

Now it is, of course, possible to form such a concept, applicable to cases of veridical perception and cases of illusion alike, and thus to siphon off the phenomenal aspects of one's visual perception. And it may further be observed that one's visual experience is, obviously, in many ways shaped by one's brain and physiology. However, none of this contradicts the common sense, direct realist view of perception. The two accounts just bypass each other, since they deploy different concepts of seeing.

When I see a tree, I do not see or experience an image (or representation) of a tree. Of course, there are only certain aspects of the tree I am (or take myself to be) aware of. (In cases of illusion, these may even be aspects the tree does not actually have.) In the line of thinking we are now considering, this partial (and possibly even distorted) visual awareness of the tree is conceptualized as an internal *object* of my visual experience. Scientists find it natural to say that this is what I *really* see, not noticing that now they have moved to a different concept of seeing. What is more, when someone who sticks to the ordinary concept of 'seeing' insists that he sees the tree itself, such scientists are prone to re-interpret this, substituting their own artificial concept of 'seeing' for the ordinary one. The result is the absurd philosophical theory that we are infallibly aware of everything that is in front of our eyes: that *all* aspects of the tree I am seeing are accurately present in my experience. Those scientists then proceed to offer an empirical refutation of this dummy theory by citing an experiment in which the subject fails to become aware of the exact shapes of the objects looked at. This, in itself,

is symptomatic of their confusion. For the philosophical theory they now oppose, under the name of 'Direct Realism', is not an empirical hypothesis to be tested by sophisticated experiments; its absurdity is perfectly obvious. Not only would it contain the claim that our vision was infallible, it would also imply that trees and houses cannot exist unperceived. It is, in short, a version of Berkeleyan idealism, with all its notorious conceptual problems.

As for (ii), the Kovács and colleagues (1996) experiment is not just a case of illusion, it also shows how our perceptual experiences are shaped by the operation of our brains: the physiological mechanisms underlying our perceptions influence them in such a way that we are more likely to pick out coherent patterns. This tendency, which helps us to discern what we are most likely to encounter, may in some cases lead to misperception. Hence, Smythies and Ramachandran (1997) write:

> The Representative Theory states that we do not see what actually is out there but what the brain computes is most probably out there. (Smythies and Ramachandran, 1997, p. 438)

Note first that the representative theory of perception discussed in philosophy as an alternative to direct realism says nothing of the kind. That you see things indirectly does not mean that you see them inaccurately, just as seeing things directly is not a guarantee of seeing them accurately (as already noted above). Neither does seeing things indirectly mean that you do *not* see them – this would be the most extreme and silly version of the representative theory, in which things in the visible world are not visible. Anyway, the statement produces a specious disjunction. It is like saying: 'We do not ever expect the things that will happen, we only expect things that we think likely to happen.' The appropriate response is that these two descriptions are not mutually exclusive: quite frequently, what we think likely to happen *is* exactly what happens. Similarly, granted that the brain has a tendency to try to clarify our perceptions in a way that may sometimes result in distortion, it does not follow that in the more common case, when the brain's influence is effective rather than counterproductive, our perceptions may not present us with what is actually out there.

In this respect, one should compare colour vision. Our eyes and the rest of our visual systems manipulate our colour perception in a way that facilitates the recognition of surface colours under changing illumination. Roughly speaking, they downplay the differences that

illumination might be expected to make on the chromatic appearance of the same object. Does that show that we do not see colours? No, on the contrary. These perceptual mechanisms increase our ability to recognize surface colours (as opposed to their varying appearances) in different light.

Now let us imagine an acoustic experiment of the same kind as the Kovács et al. (1996) experiment. Two texts A and B, each read by a distinctive voice, are cut up in this way:

Right ear: First word of A, 2nd word of B, 3rd word of A, etc.

Left ear: First word of B, 2nd word of A, 3rd word of B, etc.

Suppose we find that the subject does not hear two jumbled successions of words, but coherent texts, either A or B, depending on which one he concentrates on. Roughly speaking, the 'brain' makes the jigsaw pieces fit.

Does that show that the subject does not really hear the words that are spoken, but only representations of those words? Of course not – there is not the slightest reason to say that. The elements are all perceived just as they are; it is only their arrangement that is changed.

Similarly, in the Kovács experiment, even if we allow ourselves to say, in popular science style, that – when presented with visible objects in an extremely anomalous and confusing manner – the brain manipulates our vision, this manipulation is merely one of seeing the visible elements in a more coherent spatial ordering. There is no suggestion of our not seeing the elements that are there, but only representations of them instead. So, the experiment does not give any support to a *representative* theory of perception.

12.5 Interpreting the subjects' reports

That the reports emerging from the subjects in the Kovács experiment need some interpretation is manifest, since it is unclear which of the two concepts of seeing they are framed in terms of. In the experimental setup (as in many such setups, but also as in ophthalmology tests), the experimenters' question 'What do you see?' clearly does not invite a response using the factive concept of seeing (subjects tend not to say, 'Just let me out of this experimental device, so that I might have a better look').The experimental and the social context in which subjects' *personal* responses are being sought clearly demands the internal sense of 'seeing' here. What matters, though, is that we should not confuse this with the ordinary concept of seeing, or switch between the two.

In the experiment, since neither of the two original and complete patterns, A and B, was in the offing, our subjects may *report* that they saw these, but they can literally have done so only when we understand their reports as deploying the internal sense of 'see'. The experimental reports in question are expressions of what the subjects visually experienced, but not *true* judgements about what they *did* see in the factive sense. As experimenters 'in the know', we can report them as having seen A and B, in alternation, even though what they were looking at was, unbeknownst to them, the two strange patchwork pictures, C and D. But this clearly means only that they visually experienced (in a non-factive sense) A and B. This is, of course, the truth behind Smythies and Ramachandran's (1997) suggestion that the 'direct realist' must 'predict' that the subject will see C and D. But no *prediction* is in question here. It is quite predictable, in fact, that the experimental subjects should report seeing (in the subjective sense) no such things.

12.6 Central rivalry and bi-stable figures

Smythies and Ramachandran (1997) also remark that the Kovács and colleagues (1996) experiment, and its subjects' response to it, 'involves central rivalry similar to that shown by the Necker cube' (Smythies and Ramachandran, 1997, p. 438). (Central rivalry is the familiar phenomenon of aspect-switching, where a single stable figure is seen now as one pattern, then as another.)

This is a somewhat surprising turn (and one neither taken nor suggested by Kovács et al., 1996). Central rivalry, after all, is a well-known phenomenon, a finding which any account of perception must presumably make room for. But if Smythies and Ramachandran (1997) are correct here, *any* example of central rivalry would suffice to refute 'direct realism'. This seems unlikely. Is the 'direct realist' supposed never to have seen a Necker cube, a duck–rabbit, or any of the other examples of bi-stable figures which typically bring out central rivalry? Or is it that while the 'direct realist' may have been exposed to such phenomena, he or she has never ensured that their view of visual perception makes room for them?

In fact, in their claim that bi-stable figures lend support to representative theories of perception, Smythies and Ramachandran (1997) have things entirely the wrong way around. Following an insight from Wittgenstein, Norwood Russell Hanson pointed out long ago that bi-stable figures actually make trouble for 'sense datum' theories. To show that this is so, Hanson used the famous duck–rabbit figure. (See Figure 4.2, Hausen and ter Hark, p. 89) You might be seeing it as a duck,

while I am seeing it as a rabbit. Our visual 'sense-data' are (or can be made) identical, but our perceptual experiences nevertheless differ:

> If *a* sees the duck and *b* sees the rabbit, nothing in orthodox sense datum theory will explain this difference so long as it can be argued that their sense data are identical, geometrically indistinguishable ... So long as '*a* and *b* have the same/different visual experiences' and '*a* and *b* have the same/different visual sense data' both remain empirical claims and are not converted by the theorist into an equivalence when the argument begins to go against him, then it is meaningful to speak of situations in which two observers having indistinguishable visual sense data nonetheless have disparate visual experiences. The sense datum theory then fails completely as an analysis of visual experience. (Hanson, 1971, p. 188)

This argument will work just as well when applied to modern 'representative' theories of perception. Any supposed mental representations of a bi-stable figure (the things which Smythies and Ramachandran (1997) call 'phenomenal objects') will be just as capable of being perceived or experienced in two ways as is a physical drawing of a duck–rabbit. So, whether or not 'mental representations' or 'phenomenal objects' exist, they will not add up to what we are trying to explain, visual perception, or visual experiences, since two people could be having the same mental representations but be having different visual experiences.

What this shows is that it is inner picture theories, like the traditional version of the representative theory, that have difficulties explaining an aspect switch. We see the bi-stable figure differently, but we are not under the illusion that the figure itself has changed – as presumably we would be if our inner representation changed.

12.7 Arbitrariness

At the very end of their paper, Smythies and Ramachandran (1997) conclude that the Kovács (1996) experiment refutes 'direct realism' because

> it would obviously be most implausible to suggest that we see only what the brain computes to be probably out there when looking at C and D, but not when we are looking at A and B. Perception must depend on a brain mechanism that is unlikely to function in such an arbitrary fashion. (Smythies and Ramachandran, 1997, p. 438)

Austin would immediately have recognized this move, for he considers various related arguments. One is the indistinguishability argument, to the effect that there is no difference in kind between 'veridical' perceptions and others. Another is the argument from continuity, to the effect that one cannot put a non-arbitrary line between those perceptions in which one sees an object as having the properties it really does have, and other situations, in which one's perceptions do not have one or more of the properties one takes them to have.

No doubt, brain mechanisms cannot operate arbitrarily. But Smythies and Ramachandran are wrong to think that direct realism would be committed to the view that they do. Consider an analogous argument: Suppose you receive an email that contains the sentence: 'I shall be bcak tomorrow'. Naturally, you read 'bcak' as 'back', as there is no word 'bcak' in English, whereas 'back' makes good sense in the context, and to reverse the order of two letters is a common typo. Indeed, if you read the email very quickly, you may not even notice the mistake. So, it happens that when reading a message, what we understand to be its content is not what is actually written, but what we reckon to be probably intended. Now consider the following Smythies and Ramachandran-style argument:

> It would obviously be most implausible to suggest that we read only what we reckon to be probably intended when looking at an obvious typo, but not when we are looking at correctly typed sentences. The interpretation of written messages is unlikely to function in such an arbitrary fashion. Hence, we *always* read only what we reckon to be probably intended.

The reply is twofold: First, there is nothing arbitrary about making corrections only when the uncorrected text does not make any (good) sense. Secondly, the conclusion that we always go for the most likely reading is quite correct, but it does not mean that we do not normally read and understand exactly what is on the page or screen, since we have no reason to doubt that the text was intended as it stands.

Similarly in the case under discussion: For one thing, there is nothing arbitrary about a brain mechanism that modifies our perception only in cases where otherwise our visual impressions are utterly confusing. It is not difficult to imagine that, in such a case, neural impulses from different areas of our eyes are processed differently if that produces an impression of more familiar shapes or objects. For another thing, we can agree with Smythies and Ramachandran that certain brain mechanisms are set up in such a way that we see, or take ourselves to see, what 'is

most probably out there'. But (as already noted above) normally, what is most probably out there will be what is, in fact, out there; so that, normally, those brain mechanisms will not interfere with our seeing what is visibly in front of us.

Smythies and Ramachandran's accusation of arbitrariness is a version of the familiar Argument from Illusion (see above). If what we see when looking at C and D is not actually out there (at least not in that arrangement), but is only a construct of the brain, yet the experience of looking at C and D is phenomenologically indistinguishable from the experience of looking at A and B, it is claimed to follow that what we see when looking at A and B is not really out there either. Would it not be arbitrary to hold that in one of two phenomenologically indistinguishable cases, we see what is out there if in the other case we do not? No. Our concepts of perception are not primarily concerned with the phenomenological qualities of our perceptual experiences. What kind of object, if any, is being seen (in the factive sense) depends not only on the conscious contents of the perceptual experience, but also on what kind of object is visible to the person under the circumstances. Hence, subjectively indistinguishable experiences may, indeed, be either experiences of seeing an object or experiences of merely having an illusion of seeing that kind of object.[5]

12.8 The representative theory's latest turn: illusion theory

Smythies and Ramachandran (1997) characterize the representative theory of perception that they want to defend as saying that

> [T]he phenomenal object is a construct of the central nervous system and thus phenomenal objects are not identical to physical objects, but rather represent them, in which process a series of most complex neurocomputational mechanisms are involved that lie between the retinal image and the final construct. (Smythies and Ramachandran, 1997, p. 437)
>
> The Representative Theory states that we do not really see what actually is out there but what the brain computes is most probably out there. (Smythies and Ramachandran, 1997, p. 438)

These quotations only hint at the turn which representative theories of perception, inspired and informed by neuroscience, have taken in recent years.

Until fairly recently, the version of the representative theory of perception most familiar from philosophy, psychology, and neuroscience involved the supposition that at any one time, a person with normal vision is confronted with a complete, detailed and coherent mental representation of what is in their visual field. We might call this the *televisual version* of the representative theory, since it conceives of what is 'given' to the seeing subject as similar to the picture presented by a television screen. Because of the complexity of the neural processing to which Smythies and Ramachandran (1997) refer, representative theorists have now stampeded away from this idea. Many different kinds of experiments have convinced neuroscientists (and some philosophers) that there is simply no such coherent visual representation (see Churchland, Ramachandran, and Sejnowski, 1994; Dennett 1991). However, the idea that vision involves neural representation has *not* been given up. Instead, it has been sophisticated. In order to explain vision, different *kinds* of representations, at different levels of neural processing, are now postulated, without there being at any point in this series a single picture-like representation of the sort that earlier representative theorists wanted us to believe in.

Neuroscientists and philosophers who take this turn present various experiments which show that people are unable to report on phenomena which can be quite prominent in their visual fields. They also tend to revel in the conclusion they draw from these experiments, which is that the visual world – by which is meant here the previously-postulated complete, detailed and coherent visual representation of what is in one's visual field – is a 'grand illusion' (see Noë, 2002).

Among the major weapons in the arsenal of this new kind of representative theorist are experiments illustrating the phenomena known as *inattentional blindness* and *change-blindness*. Chris Frith's (2007) book *Making Up the Mind: How the Brain Creates Our Mental World* features a typical example of the use of such experiments.

12.9 'Inattentional blindness' and 'change-blindness'

To demonstrate 'inattentional blindness', subjects watch a videotaped action-scene, having been asked to pay attention to a specific aspect of the action (e.g. by counting the number of times players on a basketball team pass the ball to one another). After the tape has been played, they are asked whether they noticed anything unusual in the action on the tape, and many of them utterly fail to report something that counts as

such (e.g. a man in a gorilla-suit wandering across the basketball court, between the players and the camera, facing towards the camera and beating his chest like a gorilla). The phenomenon even persists when the subjects' eye-movements include fixations on the part of the screen featuring the 'gorilla'.[6]

Psychologists illustrate the related phenomenon they call 'change-blindness' using two versions of a complex scene which differ in one respect – a visually significant object is removed from the first version to make the the second (See Figure 12.2). When confronted momentarily with the first version of the scene, *then* with a uniformly grey screen, and then with the other version of the same scene (all in the device known as a tachistoscope), people typically fail to report on, and thus presumably fail to notice, the absence (or presence) of the feature in question. With respect to this example, for instance, due to Ronald Rensink, Frith says:

> What this demonstration shows is that you rapidly perceive the gist of the scene: *a military transport plane on a runway*. But you do not actually have all the details in your mind. For you to notice the change in one of these details, I have to draw your attention to it. (Frith, 2007, pp. 42–3, emphasis original)

Undoubtedly, such illustrations can show something: It *is* surprising how large and salient an aspect of a scene (here, an enormous aircraft engine) can be removed without our noticing. (However, the effect works primarily because of the interposed grey screen – if this is not present, the effect is much less marked). But the conclusions that Frith draws are entirely different and far more sweeping:

> So we have to conclude that our experience of immediate and complete awareness of the visual scene in front of us is false. There is a short delay in which the brain makes the 'unconscious inferences' by which we become aware of the gist of the scene. Furthermore, many parts of the scene remain blurred and lacking in detail. But the brain knows that the scene is not blurred and also knows that an eye movement can rapidly bring any part of the scene into vivid focus. So our experience of the visual world in rich detail is an experience of what is potentially available to us rather than what is already represented in our brain. (Frith, 2007, p. 44)

Figure 12.2 Change-blindness images, reprinted with permission of Ronald Rensink

This passage raises a number of issues, for the idea that brains make inferences (conscious or otherwise), and can be said to have knowledge can hardly be direct 'conclusions' from any such experiment.[7] Let us just concentrate on the first conclusion drawn here, though, to the effect that 'our experience of immediate and complete awareness of the visual scene in front of us is false'. This would only be mandated if ordinary perceivers took their experience of visual scenes to be both 'immediate' and 'complete'. Frith (2007) presents no evidence or argument that this is the case. Immediacy is one issue, though, and completeness another. It would be an extremely naïve and rash ordinary perceiver who thought that their visual perception of a complex scene was 'complete' in the sense that, from a strictly limited presentation of that scene, they could issue a correct and *exhaustive* report on all that it contained. One does not need 'change-blindness' illustrations to conclude that there are elements of all visual scenes (or perhaps all but the very *simplest* visual scenes) which go unnoticed by ordinary perceivers. Few people in the Western world, at least, can have made it through childhood without encountering, in comics and magazines, the sort of 'spot the difference' tasks which are primitive and static versions of these 'change-blindness' experiments.

Matters are rather different with 'immediacy', which could mean more than one thing. That Frith has something like directness in mind here is clear from his book's mission statement:

> Everything we know, whether it is about the physical or the mental world, comes to us through our brain. But our brain's connection with the physical world of objects is no more direct than our brain's connection with the mental world of ideas. By hiding from us all the unconscious inferences it makes, our brain creates the illusion that we have direct contact with objects in the physical world. (Frith, 2007, p. 17)

In the second chapter of his book, Frith (2007) contends that 'Even if all our senses are intact and our brain is functioning normally, we do not have direct access to the physical world. It may feel as if we have direct access, but this is an illusion created by our brain' (p. 40).

Frith here clearly shares with Smythies and Ramachandran (1997) the idea that something people naïvely believe, something within common sense and at the core of direct realism, is problematic and can be refuted

by experiment. But what *is* this naïve presumption? Frith thinks of it as the belief that we have 'direct contact with' or 'direct access to' things in the physical world – but what would this mean? In the ordinary sense of 'direct', our contact with, and access to, visible objects in our environment *is* direct (i.e., we do not see such things *via* seeing internal objects). So, this cannot be what is meant. What is meant seems to be a magical kind of contact or access which does not involve *anything* (any distance, any time, or even any brain events).[8] But this picture is clearly no part of common sense or of direct realism; indeed, it hardly makes sense at all. The truth is that although the scientific opponents of direct realism think they are refuting something that ordinary people are somehow committed to, the belief in question is not even one capable of meaningful articulation, and it therefore has no claim to be part of common sense.

What Frith (2007) relies on to show that our contact with the external world is not direct is Hermann von Helmholtz's argument to the effect that because nerve signals do not propagate instantaneously, but in some finite time, 'our perception of objects in the outside world is not immediate' (Frith, 2007, p. 41). Notice the slide from directness to immediacy here. Notice also that the envisaged conclusion only makes sense if the concept of perception allows for the question 'How long does perception take?' In *most* contexts, there simply is no such question. Of course, there are perfectly meaningful questions about how long it takes the visual-system to process certain kinds of information. But to think that this is the same issue is already to beg the question, by treating perception not merely as *involving*, but *as* the neural processing of information. This is not something that the critic of our ordinary way of conceiving of perception can take for granted.

Alternatively, the question might have been 'How long did it take you to perceive that object/scene?'. One can imagine ordinary perceivers making various responses to this question, but even if on some occasion one were to respond, 'I saw it *immediately*', because the concept of temporal immediacy in operation here is not a *scientific* version of that concept, the truth of this response is not imperilled by Helmholtz's revelation. (The concept of immediacy has enough room in it, as it were, for the response to remain true.) Ordinary perceivers never held the theory that nerve-impulses are propagated *instantaneously* – they typically have no views on this issue – and so their unreflective perceptual judgements are not refuted by the scientific discovery that this is not so.

12.10 Conclusion

Our conclusion is that none of these experiments have the import that modern scientific representative theorists think they do. Ironically, what they *really* imperil are previous versions of the representative theory, such as the televisual version. But its idea that in vision we are each 'given' an accurate and complete internal representation of the scene we are looking at is no part of our ordinary concept of visual perception. The experiments do not, however, imperil 'direct realism'. If one insists on building into that view the sorts of versions of the claim that visual processing is direct or instantaneous that its opponents have in mind, it must be regarded not as experimentally refuted but as already lacking sense. But the experiments do not threaten the sort of conclusions that emerge from work on the conceptual analysis of perceptual concepts: that perceptual verbs, in their primary everyday employment, are success-verbs, and that physical objects, but *not* mental representations (whether old-style or new-fangled), are among the things that people can be correctly said to see. If one *adds* to these ideas the thought that we do not see the public things we see in virtue of seeing (or even 'having') something *else*, something mental and properly called a *representation*, one has a commonsensical and defensible version of 'direct realism'.

Notes

1. See Hacker's critiques of Gregory's work (Hacker, 1991a), and of David Marr's (Hacker, 1991b), and Hyman's critique of Lawrence Weiskrantz's work on 'blindsight' (Hyman, 1991).
2. Smythies' own characterization of the visual field as 'the spatial array of visual sensations available to observation in introspectionist psychological experiments' (Smythies, 1996, p. 369) is not one that the 'direct realist' should accept, for it incorporates at least two confusions (that the contents of the visual field are *sensations*, and that sensations can be said to be observed).
3. See Quinton (1973, ch. 2).
4. See Bennett and Hacker (2003, ch. 3) and Kenny (1984).
5. See Hinton (1973); Hyman (1992); Snowdon (1990).
6. The videotape in question can easily be found by entering 'Gorillas in our midst' into an Internet search-engine.
7. See Bennett and Hacker (2003), *passim*.
8. Frith's remarks here are reminiscent of Schopenhauer's reasoning that since the objective world 'cannot just step into our heads from without, already cut and dried, through the senses and the openings of their organs', we have to accept the transcendental idealist view according to which the world is *created* by the understanding (Schopenhauer, 1847, p. 78).

References

J. L. Austin (1962) *Sense and Sensibilia* (Oxford: Oxford University Press).
M. R. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).
P. S. Churchland, V. S. Ramachandran, and T. J. Sejnowski (1994) 'A Critique of Pure Vision' in C. Koch and J. Davis (eds) *Large-Scale Neuronal Theories of the Brain* (Cambridge, Mass.: MIT Press), 23–60.
F. Crick (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul* (London: Simon and Schuster).
D. C. Dennett (1991) *Consciousness Explained* (London: Allen Lane).
C. Frith (2007) *Making Up the Mind: How the Body Creates our Mental World* (Oxford: Blackwell).
P. M. S. Hacker (1991a) 'Experimental Methods and Conceptual Confusion: An Investigation into R. L. Gregory's Theory of Perception', *Iyyun, the Jerusalem Philosophical Quarterly*, 40, 289–314.
—— (1991b) 'Seeing, Representing and Describing: An Examination of David Marr's Computational Theory of Vision' in J. Hyman (ed.) *Investigating Psychology: Sciences of the Mind After Wittgenstein* (London: Routledge), 119–54.
N. R. Hanson (1971) 'On Having the Same Visual Experiences' in N. R. Hanson (ed.) *What I Do Not Believe, and Other Essays* (Dordrecht: D. Reidel), 178–89.
J. M. Hinton (1973) *Experiences: An Inquiry into Some Ambiguities* (Oxford: Oxford University Press).
J. Hyman (1991) 'Visual Experience and Blindsight' in J. Hyman (ed.) *Investigating Psychology: Sciences of the Mind After Wittgenstein* (London: Routledge), 166–200.
—— (1992) 'The Causal Theory of Perception', *The Philosophical Quarterly*, 42, 277–96.
A. J. P. Kenny (1984) 'The Homunculus Fallacy' reprinted in A. J. P. Kenny (ed.) *The Legacy of Wittgenstein* (Oxford: Basil Blackwell), 125–36.
I. Kovács, T.V. Papathomas, M. Yang, and A. Fehér (1996) 'When the Brain Changes Its Mind: Interocular Grouping during Binocular Rivalry', *Proceedings of the National Academy of Science*, 93, 15508–11.
A. Noë (2002) *Is the Visual World a Grand Illusion?* (Thorverton: Imprint Academic).
—— (2009) *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness* (New York: Hill and Wang).
A. M. Quinton (1973) *The Nature of Things* (London: Routledge).
A. Schopenhauer (1847) *On the Fourfold Root of the Principle of Sufficient Reason,* trans. E. F. J. Payne (La Salle, Ill: Open Court Press, 1974).
J. R. Smythies (1956) *Analysis of Perception* (London: Routledge and Kegan Paul).
—— (1965) 'The Representative Theory of Perception' in J. R. Smythies (ed.) *Brain and Mind: Modern Concepts of the Nature of Mind* (London: Routledge and Kegan Paul), 241–63.
—— (1993) 'The Impact of Contemporary Neuroscience and Introspection Psychology on the Philosophy of Perception' in E. Wright (ed.) *New Representationalisms: Essays in the Philosophy of Perception* (Aldershot: Avebury), 205–31.

—— (1994) *The Walls of Plato's Cave: The Science and Philosophy of Brain, Consciousness, and Perception* (Aldershot: Avebury).
—— (1996) 'A Note on the Concept of the Visual Field in Neurology, Psychology, and Visual Neuroscience, *Perception*, 25, 369–71.
—— (2005) 'How the Brain Decides What We See', *Journal of the Royal Society of Medicine*, 98, 8–20.
—— (2009) 'Philosophy, Perception, and Neuroscience', *Perception*, 38, 638–51.
J. R. Smythies and V. S. Ramachandran (1997) 'An Empirical Refutation of the Direct Realist Theory of Perception', *Inquiry*, 40, 437–8.
P. Snowdon (1990) 'The Objects of Perceptual Experience', *Proceedings of the Aristotelian Society*, 64, 121–50.

13

Terror Management, Meaning Maintenance, and the Concept of Psychological Meaning

Timothy P. Racine and Kathleen L. Slaney

For nearly five centuries, modern science has gradually chipped away at the notion that we live in a meaningful world. We occupy a remote tiny planet in an otherwise prosaic corner of a massive and expanding universe that our species has come to dominate through a remarkably contingent and fortunate series of events. But our bodies, our species, and even the planet itself will ultimate crumble. Our death is inevitable, and we cannot know the time it will come. At best, we can hope that it will be brief and relatively painless.

It would probably be surprising to many laypeople that simply reading and thinking about the previous paragraph, and then distracting themselves from it for a short interval through some other task, might be enough to cause at least some readers to espouse, for example, more pro-nationalist or pro-religious sentiments than they otherwise might (a so-called 'worldview defence'; for reviews and a meta-analysis of this body of work, see Burke, Martens, and Faucher, 2010; Greenberg, Solomon, and Arndt, 2008; Hayes et al., 2010). The typical claim when such an effect is observed is that it reminded readers of their mortality (engendered so-called 'mortality salience'), which in turn engaged terror management (Greenberg, Pyszczynski, and Solomon, 1986; Pyszczynski, Greenberg, and Solomon, 1997) or possibly meaning maintenance mechanisms (Heine, Proulx, and Vohs, 2006; Proulx and Heine, 2006) to maintain adaptive functioning.[1] It is probably no coincidence that concerns about mortality are among what Hume called 'the ordinary affections of human life', namely 'the anxious concern for happiness, the dread of future misery, the terror of death, the thirst of revenge, the appetite for food and other

necessaries' that he took to be essential for the flourishing of human religions (Hume, 1957, p. 166).

A source of considerable justification and perhaps even pride for researchers interested in the empirical investigation of these issues seems to lie in the desire to take the often woolly concepts of Western existentialism, psychoanalytic thought, and earlier historical speculation like Hume's and run them though the lathe of a thoroughgoing experimental psychology. For example, Koole, Greenberg, and Pyszczynski (2006), advocates of what they call 'experimental existential psychology', note that:

> The human struggle with the givens of existence has captured the imagination of poets, prophets, and philosophers across the ages. Experimental psychologists are now studying people's existential concerns using the rigorous methods of psychological science. (Koole, Greenberg, and Pyszczynski., 2006, p. 215)

However, the precision of experimentation is useful to the extent that one is studying a phenomenon that is conceptually adequate, in the sense that the concepts we use to reference it are not plagued by too much conceptual ambiguity. One significant risk is that the concepts of existential philosophy or psychoanalytic thought might live in conceptual spaces that are mostly independent of those of interest to experimental psychologists. If so, such concepts can take on meanings that not only do not make the same sense when placed in their new context, but rather have no clear sense at all because their meanings are ambiguous.

In this spirit, our goal in this chapter is not to review the empirical findings that exist in this literature or to assay their methodological adequacy, but rather to examine the theoretical statements in this area with respect to their conceptual adequacy. Our goal will be to perform a case study sort of conceptual analysis of the two main explanations of these findings in the social psychological literature. We focus on the most fundamental concept, namely 'meaning', which we paraphrase as 'psychological meaning' to make it explicit that the concern of these researchers is with the experience (conscious or otherwise) of meaning, and not simply with the fact that one thing might mean another, or the like.[2]

13.1 The roots of psychological meaning

The interdisciplinary anthropologist Ernest Becker (1971, 1973) argued that much of our psychological experience is organized to avoid and

compensate for the ostensibly grim reality of our own deaths. Primarily through the use of psychoanalytic theory, Becker argues that human culture and civilization is essentially a defence mechanism writ large.

Terror management theory (TMT), which was introduced by Greenberg and colleagues (1986), was explicitly grounded in Becker's work: 'Ernest Becker, who laid the conceptual groundwork for TMT, argued that the human quest for meaning was the most basic route through which people find safety in the frightening world in which we live' (Pyszczynski, Solomon, and Greenberg, 2003, pp. 137–8). TMT lays great emphasis on the role that self-esteem, which they define as 'the feeling that one is an object of primary value in a meaningful universe', plays in reducing the anxiety associated with the terror of death. True to Becker, they claim in particular that:

> Individuals sustain self-esteem by maintaining faith in a culturally derived conception of reality (the cultural worldview) and living up to the standards of value that are prescribed by that worldview. From the perspective of terror management theory, people need self-esteem because it is the central psychological mechanism for protecting individuals from the anxiety that awareness of their vulnerability and mortality would otherwise create. (Greenberg et al., 1992, p. 913)

So-called 'self-expansive motives' are also implicated in TMT, which are said to be 'oriented toward the growth and expansion of an individual's competencies and internal representations of reality' (Pyszczynski, Greenberg, and Solomon, 1997, p. 1).[3]

Although Hayes and his colleagues (2010) report that mortality salience effects are independent of general meaning-making threats (see also Greenberg, Solomon, and Arndt, 2008), TMT and meaning-making interpretations of studies are conceptually linked through being concerned with psychological meaning in the sense discussed by Becker and others. Furthermore, TMT and its main alterative, the meaning maintenance model (MMM; Heine, Proulx, and Vohs, 2006; Proulx and Heine, 2006), draw upon Western existentialist philosophy in buttressing their empirical claims, as does Becker.[4] However, despite their rooting in existentialism through Becker, and independently of Becker, as we examine in more detail below, TMT and especially MMM explanations rely surprisingly heavily on cognitivist assumptions. That is to say, both are firmly steeped in an ontological perspective that locates all things psychological – including psychological meaning – 'in the head', in one sense or another.

Running along related, but seemingly distinct, fault lines, Jerome Bruner (1990, p. 2), an architect of the so-called 'cognitive revolution' in psychology, noted that '[The cognitive revolution] was, we thought, an all-out effort to establish meaning as the central concept of psychology – not stimuli and responses, not overtly observable behaviour, not biological drives and their transformation, but meaning.' Bruner might find the turn of phrase 'not biological drives and their transformation, but meaning' possibly apropos for this body of work because seemingly contra Bruner (1990), and also probably existentialist writers in general, the argument in TMT seems to be that a transformed 'biological drive' *is* responsible for our existential predicament. By contrast, as Bruner puts it, quoting the anthropologist Clifford Geertz, human beings are 'incomplete or unfinished animals who complete or finish ourselves though culture' (Bruner, 1990, p. 12). Bruner (1990, p. 4) also argues that, 'information is indifferent with respect to meaning' and bemoans the state of cognitivism and how it had become transformed into the study of information processing rather than meaning.[5] He sought to right these perceived wrongs by foregrounding the sociocultural constitution of human meaning that he took to be presupposed and neglected; Bruner also criticized the reductionist enterprise upon which stimulus-response and information-processing psychology were based.

In a similar vein, the meaning maintenance model (MMM; see Heine, Proulx, and Vohs, 2006; Proulx and Heine, 2006) argues that mortality-related concerns are but one application of more general concerns with meaning: '[MMM] elaborates on this existential hypothesis, proposing that human beings innately and automatically assemble mental representations of expected relations, systems that they strive to make coherent and consistent' (Proulx and Heine, 2006, p. 310). It is ironic in our view, however, that the advocates of MMM frame the introductory articles to their theory (Heine, Proulx, and Vohs, 2006; Proulx and Heine, 2006) with a description of a much earlier article by Bruner (Bruner and Postman, 1949). Heine and colleagues (2006, p. 88), for example, claim that 'according to Bruner and Postman, people maintain mental representations of expected relations, paradigms, that in turn regulate their perceptions of the world.' Although this is true to the general cognitivism in Bruner, it certainly downplays the emphasis he places on the sociocultural constitution of meaning in his later writings.

However, that the computational view of meaning and mind is still very much alive and well in psychology and philosophy despite the fact that sociocultural views have become more common is not entirely

surprising, because Bruner seems to be talking about two quite different things. On the one hand, he is making a claim about what meaning *is* (i.e., a definitional issue), whereas information-processing theorists are typically making claims about *how* agents are able to perform tasks to which knowledgeable scientists attribute certain sorts of meanings (i.e., an empirical issue). Therefore, as long as we keep in mind that language use, including information processing language use, is often metaphorical, it might still be helpful for researchers to model how it is that agents can do the things to which observers attribute meaning.

As in Bruner, and the information-processing theorists he criticizes, in TMT and MMM there are different senses of meaning used. Also, conceptual and empirical issues are sometimes conflated. Both of these possibilities are impediments to scientific progress in this research area. In the next sections of this chapter, we analyse and clarify the concept of 'psychological meaning' through performing an illustrative survey of the manner in which the concept is used in this literature.

13.2 Clarifying the meaning of 'psychological meaning'

Although TMT draws explicitly on Becker's writings, particularly Becker's (1973) *The Denial of Death*, no specific definition of 'psychological meaning' is provided in Becker's landmark work. This reflects, no doubt, that it is the ordinary, everyday use that one might call 'existential meaning' that is in play. Becker (1973, p. 6) remarks that a person 'has to feel and believe that what he is doing is truly heroic, timeless and supremely meaningful.' The key issue for Becker, as made central in TMT, is that

> Man is literally split in two: he has an awareness of his own splendid uniqueness in that he sticks out of nature with a towering majesty, and yet he goes back into the ground a few feet in order blindly and dumbly to rot and disappear forever. It is a terrifying dilemma to be in and have to live with. (Becker, 1973, p. 26)

For our purposes, Becker's everyday notion of 'psychological meaning' in his work amounts to something like a sense of purpose, significance or value, a commitment in the 'crucial projects of a person's life' (Becker, 1973, p. 215). When Becker's ideas are framed as TMT, it is claimed that 'The concept of death is, in and of itself, a type of meaning: an expectation that one's life will cease' (Pyszczynski et al., 2006, p. 335). Furthermore, as suggested by TMT:

> all cultural worldviews serve an important anxiety-reducing function by providing a sense of meaning and a recipe for attaining either symbolic or literal immortality.... Psychologically, then, the function of culture is not to illuminate the truth but rather to obscure the horrifying possibility that death entails the permanent annihilation of the self. Psychological equanimity thus requires us to believe that we live in a meaningful universe in which some form of ourselves continues to exist forever. (Pyszczynski, Solomon, and Greenberg, 2003, p. 21)

In TMT's view, there are also levels, or different senses, of meaning that need to be kept separate from one another. In responding to MMM, TMT advocates claim that:

> TMT posits that cultural worldviews function, in part, to protect people from the potential for anxiety to which this particularly upsetting meaning [i.e., death] gives rise. Thus, it is not that low level meanings exist only to quell the fear of death: meaning serves many practical and pragmatic functions involving staying alive, finding mates, caring for one's offspring, and meeting various more specific needs and goals. TMT posits that the potential for anxiety that results from the awareness of death biases the sorts of meanings people prefer, such that construals of reality that deny death in either a literal or symbolic manner are preferred. (Pyszczynski et al,. 2006, p. 335)

Therefore, it seems that according to TMT, death is a 'high level' meaning concept used to denote an expectation about mortality that, in turn, organizes lower-level meanings ('crucial projects of a person's life') that can serve both practical and existential (viz., managing death meaning) ends.

Something is amiss here, though, because when discussing death as a higher-level concept, TMT is no longer talking about a sense of purpose or engagement in some personally relevant crucial project, but rather simply an expectation about death, which is more an appreciation of death or understanding of its inevitability. Unless death itself is a personally relevant crucial project, which it may well be to Becker and TMT, this is using meaning in a different sense than the ordinary. That is, they are no longer talking about 'psychological meaning' but something else, and no clear sense (definition) has been given by TMT to this new use.

For their part, after Heine, Proulx, and Vohs (2006) and Proulx and Heine (2006) review Bruner and Postman (1949)'s study of expectations

violated through playing cards that represent, for example, a black queen of hearts, they ask, 'Why care about playing cards?' Their answer is that 'the unease experienced by participants in Bruner and Postman's (1949) study reveals a much broader concern that underlies a diverse array of human motivations. This unease reflects a need for meaning' (Heine, Proulx, and Vohs, 2006, p. 89). Although they then go on to review Western existential philosophers to better determine what this need for meaning might be, something is amiss here, too. It may well be that the absurdity of life might strike one in the moment of viewing a black queen of hearts, but there is not any obvious reason why it should. If they are to be relevant to the concept 'psychological meaning' in an ordinary sense, violated expectations would only matter existentially if the interference were to occur during a project upon which one places great personal value. The problem here is that MMM is using meaning initially in the Bruner and Postman (1949) sense as implying frustration experienced when expecting things to go one way and having them go another and then in the existential sense. Such experiences are commonplace, though, and need not cause any obvious sort of existential anxiety or threat to meaning in the way that we take to be of interest to this research community. It is akin to confusing what one might call 'informational meaning' ('Why is this happening?') with "existential meaning" ('Is this something that convinces me of my personal value or that I live in a reasonable universe?'). These uses are clearly distinct, and in our view, the senses of meaning are not related in the ways that MMM needs them to be. Therefore, it is surprising that they appeal to Western existentialism to develop their argument further.

This is not a minor issue that amounts to setting up a model with a clever study that does not quite relate to the logic of what follows; the conflation of informational and existential meaning is institutionalized in MMM:

> People are meaning-makers insofar as they seem compelled to establish mental representations of expected relations that tie together elements of their external world, elements of the self, and importantly, bind the self to the external world. When elements of perceived reality are encountered that do not seem to be part of people's existing relational structures, or that resist relational integration, these inconsistent elements provoked a 'feeling of the absurd,' a disconcerting sense of fundamental incongruity that motivates people to re-establish a sense of normalcy and coherence in their life. (Heine, Proulx, and Vohs, 2006, p. 89)

Proulx and Heine (2006) go further in this direction and attempt to link these conceptually-distinct phenomena by merely using the concept 'absurd', notoriously associated with Western existentialism, to describe the playing card stimuli used in Postman and Bruner's (1949) study:

> What if one is presented with absurd cards whose associated features violate the playing card paradigm, the existing system of expected associations that one imposes on subsequent experiences of playing cards? ... This 'feeling of the absurd' (Camus, 1955, p. 22) provoked by the meaninglessness of death does not differ in kind (although, it surely differs in magnitude) from the meaninglessness elicited by a black queen of diamonds. (Proulx and Heine, 2006, pp. 309–10)

However, if an ordinary sense of existential meaning is in play, then these forms of meaninglessness *do* differ in kind.[6] Nonetheless, Proulx and Heine (2006) feel justified in arguing that Bruner and Postman (1949) should be placed on a list of 'the most important works yet published on meaning', within which they include theorists such as Frankl, Freud, Maslow, Rogers, and Becker (1973).[7] To be clear, we are not opposed to bringing existential philosophy, psychoanalytic, or Rogerian approaches and the like into novel forms of psychological theorizing, but this characterization of existentialism, like that of TMT's, seems to miss out on much of what motivated this form of philosophy. Essentially, despite the fact that Proulx and Heine (2006, p. 310) take Kierkegaard's discussion of absurdity to centre around 'unexpected associations', in our view this is schema talk mixed with existential speak.[8] Furthermore, as in their use of Bruner and Postman (1949), the concept 'mental representation' does not come into the discussion for existentialist philosophy, to our knowledge. If it were to, it would most likely be in a critical spirit in our view.

13.3 Logical and empirical issues in psychological meaning

One thing that should be apparent from the previous section is that it would be difficult to compare the ideas of TMT and MMM if they rely on different meanings of 'psychological meaning'. This definitional issue is independent from the empirical issues of how measures hang together or even the interpretation of meta-analyses. That is, one would not conduct a factor analysis or meta-analysis to resolve a definitional issue, as such empirical analyses presuppose a certain amount of conceptual

clarity; what would be required is analysis of the concepts themselves (i.e., conceptual analysis).

As an illustration of the distinction between empirical and definitional issues, let us begin by considering Proulx and Heine's claim that:

> whatever meaning is, it must be a broad, practically all-encompassing psychological construct. Actually, this is very much the case, and it should therefore be of little surprise that it is already well entrenched in the psychological literature, ubiquitous across disciplines, albeit hidden within the current psychological nomenclature. Whatever it happens to be called, *meaning* means the same thing: mental representations of expected relationships. (Proulx and Heine, 2006, p. 310, their emphasis)

Although the foregoing sounds like a series of empirical claims, this is misleading. The view in Proulx and Heine (2006) is that (a) meaning is 'mental representations of expected relationships', which is simply a definition of meaning – and a rather unhelpful one given the phenomena of concern – that they, mistakenly in our view, attribute to existentialists, and (b) they and others have discovered that meaning is a 'broad, practically all-encompassing psychological construct.'

As regards (a), when people speak of meaning/meaninglessness, they are not talking about their mental representations but often about a moment of existential crisis or some crisis of faith or the like. What has happened here is that Proulx and Heine (2006) have taken what they consider to be the empirical *cause* of a variety of behaviours and used this as a *definition* of those behaviours. Further, there is no empirical evidence that anything of this sort is occurring; this is simply how a cognitivist describes (or, as the case may be, explains) behaviour. But, is it a reasonable move to make in this case? That is, what would be the grounds for saying that an organism has mentally represented some expected relation? Well, *what* expected relation are we talking about here? If one is thinking about the untimely death of her father in the right sort of circumstance and manner, she might be said to be struggling to understand its meaning, and this might well make her own mortality salient. Where does the mental representation come in? Well, let us imagine that she says she cannot get the upsetting image out of her mind; we would know what she means, and can understand why one might describe her being under the grip of a troubling mental representation. This is what the cognitivist, be they MMM or TMT, has in mind. They mean to *stipulate* that a certain class of behaviour always involves

a mental representation. The trouble, and why such claims bear a non-empirical character, is simply their stipulative nature. No effort is made to consider whether a particular case may or may not involve representation in an ordinary sense of the concept. And critically, it is not part of the ordinary use of the concept of 'psychological meaning' that representation is necessarily (and therefore definitionally) involved. What is essential is that some purpose and commitment in crucial projects of a person's life be involved; that is how the term is defined, at least in the ordinary sense. If one likes, given that mental representations may or may not be involved, by definition they are not an essential part of the concept. Furthermore, even if mental representations *were* always involved, it still does not follow that having a mental representation is a definitional criterion for 'psychological meaning.'

As regards (b), the empirical investigation of (existential) meaning presupposes that a definition, which we have suggested to be a fairly mundane one, is already in place. To find out through empirical means that persons deal with existential issues in particular ways or in particular circumstances has no necessary bearing on the meaning of this term. In such cases, one might have discovered something about the phenomenon of psychological meaning, but they would not have discovered anything about the *concept* that denotes it, as to count as a discovery *about* psychological meaning presupposes that 'psychological meaning' can be defined. This does not mean that empirical discoveries cannot *motivate* conceptual change, rather, simply, that, in this case, the concept of 'psychological meaning' must already be defined if any empirical claims made about it are to have any sense.[9] If one adjusts or *thinks that one has* adjusted or otherwise discovered the meaning of a concept that one is 'studying' empirically, which seems to be the case here given the description of 'whatever meaning is', one has overlooked the fundamental distinction between empirical and definitional issues.

Although this is not the time or place for a thorough-going analysis of the scope and limits of cognitivism, there is an immense critical literature on this topic that does not seem to have come to the attention of advocates of TMT or MMM. The problem is that the adequacy of these theories rests largely on an array of contestable assumptions that are neither justified nor barely discussed, and whatever the empirical yield might be of this body of work, a house is only as solid as its foundations.

Pyszczynski, Greenberg, and Solomon (1997, p. 2), however, may do Proulx and Heine (2006) one better by claiming that TMT 'requires only one commonly accepted and rather uncontroversial *a priori* assumption: specifically, that living organisms are oriented toward self-preservation.'

It is a tribute to the uncritical acceptance of a variety of contestable cognitivist assumptions and metaphors in modern academic discourse that the following statement can be said to contain only one (commonly accepted) assumption:

> The theory posits that this terror is managed by a dual-component cultural anxiety buffer, consisting of (a) an individual's personalized version of the cultural worldview, which consists of a set of benign concepts for understanding the world and one's place in it, a set of standards through which one can attain a sense of personal value, and the promise of literal and/or symbolic immortality to those who live up to these standards; and (b) self-esteem, or a self of personal value, which is attained by believing that one is living up to the standards of value that are part of the cultural worldview. Because of the role these structures play in controlling anxiety, a great deal of energy is devoted to their maintenance and defence. (Pyszczynski, Greenberg, and Solomon, 1997, p. 2)

Although those familiar with the relatively large number of studies cited in support of TMT may be shocked by our apparent naiveté, the above is not a set of empirical claims, and that therefore no amount of evidence could support (or refute) them. What these studies show is that when people are prompted to become aware of things that remind them of death, they tend to, for example, engage in world-view defences.[10] However, the *a priori* assumptions in the above are numerous and include: (a) there is a mechanism (specifically a buffer) that is responsible for terror management; (b) it contains mental representations ('personalized versions') of the cultural worldview in a set of concepts; (c) it contains representations of standards that one can follow to experience personal value; (d) humans are in a state of near constant anxiety at a subconscious or unconscious level about their mortality; and, (e) this anxiety reduction is managed by maintaining said mechanism. Parts of this proposal are assumptions that TMT shares with Becker (1973) concerning chronic anxiety related to mortality and the extent to which it may be responsible for adaptive behaviour, but other assumptions are attributable to the dominance of the representational theory of mind (RTM) that holds sway in much of psychology and philosophy. Whatever the limits of RTM, which we believe to be considerable when applied to the wrong sort of phenomena, at the very least this shows some invisible philosophizing on the part of TMT theorists.

We have suggested in previous sections that cognitivism is the common framework shared by TMT and MMM. We now advance this claim in more detail and suggest that it is actually the wrong sort of approach to take to phenomena involving existential meaning.

13.4 TMT, MMM, and RTM

Most readers, including TMT and MMM theorists, will be acutely familiar with the cognitivist framework ushered in by the cognitive revolution in the late 1950s. What is sometimes less commonly understood is that cognitivism is a theory, or perhaps set of theories, often characterized and defended in the philosophy of mind under the banner of the representational theory of mind (RTM). The earliest formal descriptions of RTM were elaborated by the British empiricists (Slaney and Racine, 2011). In brief, the most basic idea is that mental intermediaries symbolically represent to agents the objects or states of affairs that they perceive in the external world. These relations are claimed to be mediated by concepts that are taken to be mental entities with semantic content; concepts are claimed to be distinguished by differences in mental representational structure.

We suspect that our attempt to expose and problematize the cognitivist roots of TMT and MMM will not necessarily lead to an immediate laying down of arms from the researchers in this community. However, our issue with TMT and MMM and indeed RTM-styled explanations mainly lies in the fact that such theories and models are radically underspecified and not scientifically very useful. Although this is by no means particular to this body of literature, essentially they amount to redescriptions of the phenomenon with a mentalistic gloss. For example, despite the popularity of these sorts of terms of art, we would argue that TMT advocates have not really told us anything additional when they speak of a 'dual-component cultural anxiety buffer.' The terror management mechanism, in our view, is just a metaphorical way of describing how and perhaps why agents respond to mortality threats in the way that they do. To move beyond metaphor would require that one could actually lay out this claim in an empirically-tractable manner. This would also require that all instances of terror management involve mental representation in a non-stipulative sense (i.e., by showing that all phenomena that fall under this concept intrinsically involve mental representation). However, this is a logical impossibility because it is, on pain of repetition, upholding purposes in life that defines this concept, and this need not involve mental representation at all.

As for MMM, recall that Heine, Proulx, and Vohs (2006, p. 88) interpret Bruner and Postman (1949) as claiming that 'people maintain mental representations of expected relations, paradigms, that in turn regulate their perceptions of the world.' This is despite the fact that, as in the case of the Western existential literature, nowhere in Bruner and Postman's article is the term 'mental representation' used. It is similarly a tribute to the hegemony of cognitivism that theorists can take a study where a concept has not really been invented in its attributed form – recall that the Cognitive Revolution took hold in the late 1950s – and ascribe it to theorists who do not yet hold it. As noted, in the case of Bruner and Postman (1949), this is forgivable in the sense that this is generally consistent with the cognitivist description of mental functioning that Bruner and others led many psychologists to embrace until this day. However, with the proliferation of embodied and distributed cognitive models, and even talk of 'extended minds', there is reason to think that cognitivism is beginning to lose its hold on some (Susswein and Racine, 2009). Curiously, these post-cognitivist forms of theorizing have yet to make their way into this body of work, which itself seems to beg for a non-RTM explanation.[11] Although we can overlook MMM's description of Bruner and Postman, MMM's projection of RTM-styled mechanisms onto the writings of Western existential philosophers is less conscionable.

However, Proulx and Heine (2006, p. 310) claim to elaborate the 'existential hypothesis, proposing that human beings innately and automatically assemble mental representations of expected relations, systems that they strive to make coherent and consistent.' Setting aside the fact this existential hypothesis is neither true to existentialism nor a hypothesis, this is another RTM-styled redescription of a set of empirical findings. As with TMT, all one can really say is that this is a specification or perhaps an illustration of a way of thinking about how people function with respect to their knowledge of their mortality. However, again, the use of RTM does not add anything new here. All it really manages to do is provide a certain form of vocabulary in which to couch descriptions of the phenomena under study, but one that is, unfortunately, quite ill-suited, given the massive generality of such phenomena.

The key issue is that if the aim of TTM and MMM is to provide theories (explanations) of psychological meaning – its nature, its properties, its relations to other phenomena – they cannot do so through a sole reliance on RTM. Such descriptions presuppose the existence and nature of mental representations; thus, when theories, such as

TTM and MMM, assume RTM, they no longer permit the deduction of the empirical propositions which can then be put to the critical (i.e., empirical) test. Rather, they recapitulate an old, albeit in many respects intuitively pleasing, picture of the mind and its contents. However, what new do we learn about how persons manage psychological meaning by redescribing it in terms like 'mental representations of expected relations' or 'dual-component cultural anxiety buffers'? In our view, very little.

13.5 Concluding remarks

We have argued that the social psychological research concerning human existential meaning has a number of serious conceptual limitations and shortcomings. There is equivocation over the meaning of the concept 'psychological meaning', which creates comparative tensions both within and between TMT and MMM, and there is conflation of empirical and definitional issues, leading to the erroneous view that we do not really understand 'psychological meaning' but can come to a fuller understanding with more empirical discovery. Because of the foregoing, it is not at all obvious that TMT, and especially MMM, can tell us anything about existential meaning, because they are not investigating meaning in an ordinary sense. In our view, these multiple shortcomings have led to theoretical accounts that are little more than RTM-styled redescriptions of the phenomena of interest.

If we have probably been a bit harder on MMM than TMT in this chapter, it is not by accident. Despite some slippage when considering levels of meaning, TMT is, in some ways, a 'better' theory because it mostly sticks to the familiar everyday notion of 'psychological meaning'. Further, it avoids the artificiality of adding Bruner and Postman, and generally stays truer to Becker and his take on existentialism. Equally important, though, is that the architects of TMT have explicitly distanced themselves from the excesses of the cognitive revolution (Solomon, Greenberg, and Pyszczynski, 2004). However, in our view, their construal of motivation bears many vestiges of cognitivism, both in structure and content. But to their credit, when they claim they were informed that TMT 'was interesting and may even have merit [but] would never gain creditability in the field without empirical support', their response was that they 'had no intention of pursuing research on Becker's ideas [because to them] they stood on their own because they helped explain much of what they knew about

human behavior' (Solomon, Greenberg, and Pyszczynski, 2004, p. 15). Now, clearly they went on to 'test' these ideas, but it would seem that it may have been with an awareness that they were illustrating Becker's way of thinking about human behaviour rather than engaging in a series of hypothesis tests.

Nonetheless, there is something deeply ironic about the fact that the empirical demonstration of the effects that the human existential conditional has upon how people live their lives has lead to a group of theories about the fact that they so do. We would guess that existential philosophers and even psychologists might well chafe at the denial of death that these theories themselves embody. As noted, although Becker's arguments are cast in largely psychoanalytic terms, one significant impetus for Becker is the work of the existential philosophers taken up by MMM as inspiration for their model. Similar to Heine, Proulx, and Vohs's (2006) MMM, Koole, Greenberg, and Pyszczynski (2006), which includes key members of the TMT group, predicate what they called experimental existential psychology mostly on Becker and the same existentialist body of work. However, it should be noted that Becker's emphasis, like the existentialists upon whom he draws, is on *maintaining* the confrontation with the absurd and not abstracting out of it through clinging to some cultural worldview. In this sense, TMT and MMM seem at odds with Becker's most central message as is suggested in this stirring passage near the end of his opus:

> The problem with all the scientific manipulators [here Becker means behaviourists such as Skinner and Watson but his concern obviously applies more broadly] is that somehow they don't take life seriously enough; in this sense, all science is 'bourgeois,' an affair of bureaucrats. I think that taking life seriously means something such as this: that whatever man does on this planet has to be done in the lived truth of the terror of creation, of the grotesque, of the rumble of panic under everything. Otherwise it is false. (Becker, 1973, pp. 283–4)

In a sense then, whereas creating theories about terror management or meaning maintenance are undoubtedly ways of boosting self-esteem through engaging in culturally-sanctioned professional activities, such activities are in the double bind of both pointing out that humans tend to do this and simultaneously abstracting out of the 'confrontation with the absurd. Certainly, TMT has done much to popularize Becker's views, but were Becker still alive, he might not have entirely approved of the project.

Notes

1. Although it is too early in the MMM research program for a meta-analysis and review to occur, Proulx and Heine (e.g. 2008, 2009, 2011) have a number of studies that report empirical findings related to MMM. Although they publish later theoretical statements (e.g. Proulx and Heine, 2010), the character of their model is captured in the original Heine, Proulx, and Vohs (2006) and Proulx and Heine (2006) that we use in our analysis. We have also stuck to a small number of key theoretical statements in our analysis of TMT.
2. Although the theorists whom we discuss typically speak simply of 'meaning' in general, their concern is more with what one might call the 'meaning of life.'
3. We should note that this is a rare case where TMT theorists explicitly evoke mental (internal) representations in their work. However, we argue that the general character of their theorizing is quite standardly cognitivist.
4. There are substantial differences within the philosophers that are typically taken to constitute Western existentialism and treating this work as monolithically relevant for 'psychological meaning' in the way that TMT and MMM does seems risky. These differences run as deep, for example, as the relation of absurdity and meaningless, and whether life is, in some sense, meaningless or not.
5. Despite couching their explanation largely within a cognitivist framework, we are aware the TMT advocates would share in such a criticism of information processing (e.g. Solomon, Greenberg, and Pyszczynski, 2004). However, the tension between TMT and Bruner (1990) stands with respect to their emphasis on biological drives and meaning. Further, although this is clearly a sensitive issue for TMT (Kirkpatrick and Navarrete, 2006; Laudan et al., 2007; Pyszczynski et al., 2006), for reasons quite different than those listed by Kirkpatrick and Navarrete (2006), in our view their use of Evolutionary Psychology-styled use of evolutionary theory *is* out of step with recent innovations in theory and research in the biological sciences (see e.g. Pigliucci, 2009), and Bruner's biological drive statement may anticipate this as well. However, this is neither the time nor place for such a discussion (see e.g. Racine, in press; Wereha and Racine, 2012).
6. Although their argument is not couched in conceptual analytic terms, TMT advocates also come to the conclusion that 'there is something fundamentally different about the two types of meaning threats in MMM' (Pyszczynski et al., 2006, p. 335).
7. They seem to distance themselves from their early exuberance for this study and discuss it as concerning paradigms and *not* meaning in later work (e.g. Proulx and Heine, 2010). They also attribute their interpretation of Bruner and Postman (1949) to Kuhn's treatment of this study in Proulx and Heine (2008). It would seem, therefore, that they eventually recognized the tension in using it in the manner that they originally did.
8. This is an eccentric reading of a concept that Kierkegaard (1976) uses to distinguish knowledge and faith. Furthermore, it is not some run-of-the-mill set of unexpected associations that is of concern.
9. For example, if one were to adopt the definition 'scores exceeding X on psychological meaningfulness measure, Y' on the basis of consistent and replicable factor analytic findings. However, it is unlikely that such a

constrained definition would ever be seriously entertained as a *replacement* for the ordinary sense of purpose and commitment in crucial projects of a person's life. Rather, the motivation for forming such technical definitions of 'constructs' tends to be strictly an epistemological manoeuvre (Lovasz and Slaney, 2013; Slaney and Racine, 2013).

10. Manipulations of unconscious or preconscious death awareness are also performed in both TMT and MMM, but they are based on the same construals of meaning and mind.
11. We do not presume that such post-cognitivist approaches are immune to conceptual problems, only that they circumvent a number of the conceptual failings inherent to RTM-based cognitivism.

References

E. Becker (1971) *The Birth and Death of Meaning* (New York: Free Press).

—— (1973) *The Denial of Death* (New York: Free Press).

J. S. Bruner (1990) *Acts of Meaning: Four Lectures on Mind and Culture* (Cambridge, Mass.: Harvard University Press).

J. S. Bruner and L. Postman (1949) 'On the Perception of Incongruity: A Paradigm', *Journal of Personality*, 18, 206–23.

B. L. Burke, A. Martens, and E. H. Faucher (2010) 'Two Decades of Terror Management Theory: A Meta-Analysis of Mortality Salience Research', *Personality and Social Psychology Review*, 14, 155–95.

J. Greenberg, T. Pyszczynski, and S. Solomon (1986) 'The Causes and Consequences of a Need for Self-Esteem: A Terror Management Perspective' in R. F. Baumeister (ed.) *Public Self and Private Self* (New York: Springer-Verlag), 189–212.

J. Greenberg, S. Solomon, and J. Arndt (2008) 'A Uniquely Human Motivation: Terror Management' in J. Shah and W. Gardner (eds) *Handbook of Motivation Science* (New York: Guilford Press), 113–34.

J. Greenberg, S. Solomon, T. Pyszczynski, A. Rosenblatt, J. Burling, D. Lyon, L. Simon, and E. Pinel (1992) 'Why Do People Need Self-Esteem? Converging Evidence That Self-Esteem Serves an Anxiety-Buffering Function', *Journal of Personality and Social Psychology*, 6, 913–22.

J. Hayes, J. Schimel, J. Arndt, and E. H. Faucher (2010) 'A Theoretical and Empirical Review of the Death-Thought Accessibility Concept in Terror Management Research', *Psychological Bulletin*, 136, 699–739.

S. J. Heine, T. Proulx, and K. Vohs (2006) 'The Meaning Maintenance Model: On the Coherence of Social Motivations', *Personality and Social Psychology Review*, 10, 88–110.

D. Hume (1957) *The Natural History of Religion*, H. E. Root (ed.) (Stanford, Calif.: Stanford University Press).

S. Kierkegaard, S. (1976) *Kierkegaard's Journals and Papers*, Vol. 1, H. V. Hong and E. H. Hong (eds) (Bloomington: Indiana University Press.)

L. A. Kirkpatrick and C. D. Navarrete (2006) 'Reports of My Death Anxiety Have Been Greatly Exaggerated: A Critique of Terror Management Theory from an Evolutionary Perspective', *Psychological Inquiry*, 17, 288–98.

S. L. Koole, J. Greenberg, and T. Pyszczynski (2006) 'Introducing Science to the Psychology of the Soul: Experimental Existential Psychology', *Current Directions in Psychological Science*, 15, 212–6.

M. J. Laudan, S. Solomon, T. Pyszczynski, and J. Greenberg (2007) 'On the Compatibility of Terror Management Theory and Perspectives on Human Evolution', *Evolutionary Psychology*, 5, 476–519.

N. Lovasz and K. L. Slaney (2013) 'What Makes a Hypothetical Construct "Hypothetical"? Tracing the Origins and Uses of the Hypothetical Construct Concept in Psychological Science', *New Ideas in Psychology*, 31, 22–31.

M. Pigliucci (2009) 'An Extended Synthesis for Evolutionary Biology', *Annals of the New York Academy of Sciences*, 1168, 218–28.

T. Proulx and S. J. Heine (2006) 'Death and Black Diamonds: Meaning, Mortality, and the Meaning Maintenance Model', *Psychological Inquiry*, 17, 309–18.

T. Proulx and S. J. Heine (2008) 'The Case of the Transmogrifying Experimenter: Affirmation of a Moral Schema Following Implicit Change Detection', *Psychological Science*, 19, 1294–300.

T. Proulx and S. J. Heine (2009) 'Connections to Kafka: Exposure to Meaning Threats Improves Implicit Learning of an Artificial Grammar', *Psychological Science*, 20, 1125–31.

T. Proulx and S. J. Heine (2010) 'The Frog in Kierkegaard's Beer: Fluid Meaning in the Threat-Compensation Literature', *Social and Personality Psychology Compass*, 4, 889–905.

T. Proulx and S. J. Heine (2011) 'Turn-Frogs and Careful-Sweaters: Non-Conscious Perception of Incongruous Word Pairings Provokes Fluid Comprehension', *Journal of Experimental Social Psychology*, 47, 246–9.

T. Pyszczynski, T. Greenberg, and S. Solomon (1997) 'Why Do We Need What We Need? A Terror Management Perspective on the Roots of Human Social Motivation', *Psychological Inquiry*, 8, 1–20.

T. Pyszczynski, T. Greenberg, S. Solomon, and M. Maxfield (2006) 'On the Unique Psychological Import of the Human Awareness of Mortality: Theme and Variations', *Psychological Inquiry*, 17, 328–56.

T. Pyszczynski, S. Solomon, and J. Greenberg (2003) *In the Wake of 9/11: The Psychology of Terror* (Washington, D.C.: APA Press).

T. P. Racine (in press) 'How Useful Are the Concepts "Innate" and "Adaptation" for Explaining Human Development?' *Human Development*, 56, 141–6.

K. L. Slaney and T. P. Racine (2011) 'On the Ambiguity of Concept Use in Psychology: Is the Concept "Concept" a Useful Concept?' *Journal of Theoretical and Philosophical Psychology*, 31, 73–89.

—— (2013) 'What's in a Name? Psychology's Ever-Evasive Construct', *New Ideas in Psychology*, 31, 4–12.

S. Solomon, T. Greenberg, and T. Pyszczynski (2004) 'The Cultural Animal: Twenty Years of Terror Management Theory and Research' in J. Greenberg, S. L. Koole, and T. Pyszczynski (eds) *Handbook of Experimental Existential Psychology* (New York: Guilford Press), 13–34.

N. Susswein and T. P. Racine (2009) 'Wittgenstein and Not-Just-in-the-Head Cognition', *New Ideas in Psychology*, 27, 184–96.

T. J. Wereha and T. P. Racine (2012) 'Evolution, Development and Human Social Cognition', *Review of Philosophy and Psychology*, 3, 559–79.

14

A Conceptual Investigation of Inferences Drawn from Infant Habituation Research

Michael A. Tissaw

Most of the time, we easily infer from young infants' behaviours that they are happy, hungry, tired or experiencing teething pain or that they saw the bird land on the windowsill, or many other things. But there are times when it is not so easy to tell how it is with them; for example, when we wonder whether crankiness and crying are due to an upset stomach, headache or impending illness. Whatever the circumstances, to attend to our infants' needs, we need to bridge an epistemological gap of sorts; a gap between behaviour and our perspective on what the behaviour means. Language is the primary tool we use in our efforts to bridge the gap. Experimental research on infant cognition is similar to our everyday interactions with infants in this fundamental respect. But, of course, in experimental contexts, researchers control variables, know in some detail what they expect to observe, and typically frame what they observe by theory. For many decades, theorizing on human infant cognition has included construction of models of the 'inner' machinery of cognition that correspond to what is observed in controlled experiments. Sometimes, researchers have invented constructs for the purpose of bridging the epistemological gap. But those constructs always have been parasitic on our ordinary ways of speaking, as illustrated in researchers' need to use ordinary concepts to explain the meanings of the constructs. Whether using ordinary words or specialized constructs, researchers have a choice on which words to choose. Assuming that the experimental methodology is fit to the purpose, that the research is well-designed, or the measurements are accurate, which words do researchers choose to describe what infants are capable of doing cognitively, and which words to they choose to characterize the machinery behind the doing?

My interests in experimental research and theorizing on human infant cognition are driven by its conceptual messiness. Thus, I am inclined to consider the question of which words to choose in light of the relationship between human developmental science and the philosophical analysis of language. In theorizing on early cognitive development, there persists a strong desire to inform or subvert centuries of philosophical debate over nature-nurture. On this issue, I admit there are good reasons for developmental scientists to try to show that science trumps 'mere philosophical speculation.' There are exceptions. But for the most part, when it comes to early cognitive development, traditional philosophy – as it is known by the scientific research community – has done little more than frame topics of interest, suggest some ways to think about human development and supply some concepts. Many of traditional philosophy's factual claims either have not weathered well or are considered mundane. No wonder developmental science should have nothing to do with philosophy! Rather, it should rely on experimental methods thought suited for getting into the minds of babes. But do the methods guarantee right decisions on which words to choose?

Psychologists of all stripes tend to stick to their words. So, I am not holding my breath in hopes of converting those I believe have chosen the wrong words. But I feel I have a duty to try. For those who are willing to listen, I ask: What if a certain kind of philosophy, which makes no claims to the facts, can tell us which words *not to choose*? That might be helpful, assuming that knowing which words not to choose can provide insight on which words to choose. In this chapter, I suggest ways in which conceptual analysis, undertaken in the spirit of Wittgenstein's (e.g. 1953) writings, can be helpful in this effort. My aim is to examine the conceptual coherence of inferences drawn in experimental research employing the visual habituation method and some of its variants.

Section 14.1 begins with a summary of the basic nature of the research, including some aspects of its historical development particularly associated with the issue of descriptive and explanatory parsimony. There I only want to suggest some reasons why we have arrived at a point where researchers routinely attribute quite sophisticated cognitive abilities to young infants. (By 'young infants', I mean infants less than a year old.) The extensive experimental literature on infant visual habituation spans approximately 50 years and continues to grow rapidly.[1] So, for purposes of expediency, in Section 14.2 I will focus on a specific case of theorizing on infants' so-called 'physical-reasoning system' by the well-known developmental researcher Reneé Baillargeon. I have chosen this case because it serves the purposes of being accessible to readers

and generalizable to many other cases of theorizing, is suitable for analysis, and is located in a review article that begins with some statements about the role that philosophy has played in the nature-nurture debate. I conclude the section by discussing a few parameters of the conceptual investigation to follow and consider some extant objections from the research community to the sort of theorizing in question. My aim in Section 14.3 is to discuss two interrelated sources of confusion in Baillargeon's theorizing. They are: (1) conflating natural abilities and acquired abilities, and (2) theorizing young human infants as persons. I discuss some implications of these sources of confusion in a brief, concluding section.

I should make a few qualifications. First, I want to be understood by members of the psychological research community. So throughout, I keep philosophical terminology – even reference to Wittgenstein – to a minimum and do not expound details on the philosophical method that informs this chapter. Second, philosophers and philosophically-minded psychologists will note that my target in this chapter is Cartesianism in human developmental science. However, I make no mention of Cartesianism hereafter. Finally, I paint in broad strokes, and what I call a 'conceptual investigation' lacks the usual detail and is far from complete. But it may suffice to get some researchers to think about other ways to describe the cognitive abilities of young infants, given their responses to variable manipulations in the laboratory.

14.1 Habituation – dishabituation

An infant is said to 'habituate' when its responsiveness to a stimulus decreases to some pre-established criterion. 'Dishabituation' is said to occur when, after a change in the stimulus, there is recovery of the habituated response (see Kahn-D'Angelo, 1987, pp. 42–3). Visual habituation experiments often employ 'looking-time' as the dependent variable. Here is a simple example I use in my undergraduate developmental courses.[2] Slater, Morison and Somers (1988) showed 16 neonates a black-and-white striped disk, oriented 45° from vertical. On the first trial, looking times averaged over 40 seconds. This average dropped below 10 seconds by the eighth trial, indicating that the neonates had habituated to the stimulus. Then, when the disk was rotated 45° from vertical in the other direction, dishabituation occurred.

Researchers have options if they choose to describe the *experiences and/or cognitive abilities* of infants during habituation and dishabituation. For example, during habituation trials, infants can be said to

be 'interested' or 'attentive' during initial exposure to the stimulus, then eventually 'not as interested' or 'bored' with it. During dishabituation, infants can be said to 'regain interest,' 'notice a difference' in the stimulus or 'be surprised'. The words used by researchers depend, in part, on the purposes of the research. Oftentimes, precedent in the literature all but makes the decision. Slater, Morison, and Somers simply wanted to suggest that some degree of visual cortical functioning is present at birth. They did not need to use psychological predicates to describe what their neonates were experiencing during habituation – dishabituation.

Actually, the purposes, experimental design and explanatory parsimony of Slater, Morison, and Somers are characteristic of early use of the habituation paradigm, when infants' responses were described in terms of physiological bases.[3] Colombo and Mitchell (2009) note that the first experiments using habituation as a technique – prior to when the term was introduced by Cohen (1966) – were aimed primarily to test visual discrimination.[4] They add that the early literature on infant visual habituation often included constructs similar to 'attention' (e.g. 'arousal' and 'orienting'). Even with the rise of cognitive psychology, researchers tried to meet the challenge of operationally defining such constructs to maximize descriptive and explanatory parsimony, while minimizing reference to 'inner' psychological experience, processes and states. But, for example, the mere labelling of Fantz's (1964) 'preferential looking method' carried psychological baggage.

During the 1970s and early 1980s, emphasis on use of traditional habituation and its variants (e.g. high-amplitude sucking and violation of expectancy) shifted from attempts to demonstrate human infants' cognitive and intellectual *competence*, to documentation of 'the extant *cognitive abilities, skills, and products* [emphasis added] possessed by the infant' (Colombo and Mitchell, 2009, p. 226). These emphases resulted in a substantial number of remarkable claims about the cognitive abilities of infants, potentially still to be weaned and months from uttering their first words. For example, from a handful of relatively recent publications we find: preferring, identifying, distinguishing, expecting, being surprised, believing, strategizing, judging, understanding, reasoning, interpreting, categorizing, knowing and explaining. As we will see, sometimes attributions of this sort have been made toward suggesting that some cognitive abilities appear prior to others, that there are discernable, piecemeal developmental courses of cognitive ability acquisition, and that development of some abilities is based on innate 'concepts,' 'categories' or 'principles.'

Researchers have never lost sight of the need to explain the physiological bases of infant's responses. Whether these bases are innate or not, identifying the underlying mechanisms has always been the goal. Currently, in the vast majority of cases, thoroughgoing explanation still must wait for experimental results to be integrated with the promises of the neurosciences. Meantime, terms such as 'physical representation' have had to suffice. In turning to Baillargeon's theorizing, we will see just about all of the above in play.

14.2 Reneé Baillargeon on young infants' physical 'reasoning'

'Innate ideas revisited: For a principle of persistence in infants' physical-reasoning' (Baillargeon, 2008) begins with a brief summary of the philosophical debate over innate ideas; from proponents Plato, Descartes and Leibniz – curiously, there is no mention of Kant – to opponents Locke, Hume and Mill, then Watson, Skinner and Piaget. This little history lesson concludes with endorsement of Chomsky's (1965) universal grammar which, says Baillargeon, differs from earlier rationalist views on two counts. First, the universal grammar is 'an unconscious language-acquisition system' and not 'a set of ideas that can be brought into consciousness by appropriate triggers.' Second, 'the system is construed as a biological adaptation whose existence is rooted in the process of evolution, rather than in metaphysics' (Baillargeon, 2008, p. 2). So, an unconscious system replaces a set of innate ideas and evolutionary biology replaces metaphysics.[5]

Baillargeon is one of many cognitive scientists who have adopted Chomsky's views in thinking that one major task of developmental science is to investigate the possibility of innate systems foundational to some cognitive abilities. In principle, there is nothing wrong with the proposal to associate cognitive abilities of young infants, thought to be evident in their behaviours, with underlying neurological systems. If the abilities are in evidence early enough, there may be reason to conclude (provisionally) that the system is innate. There are variations on depth of the theorizing. Some researchers (e.g. Wilcox, 1999) associate their experimental results with specific and somewhat detailed evidence from neuroscience.[6] We will see that Baillargeon does so obliquely by employing such terms as 'physical representation' and 'physical-reasoning system.' But her primary aim is to show that the principles of *continuity* and *cohesion* which, according to Elizabeth Spelke (e.g. 1988), guide infants' interpretations of physical events and 'represent only two

corollaries of a single and more powerful principle of *persistence*, which states that objects persist, as they are, in time and space' (Baillargeon, 2008, pp. 2–3).

My concerns lie not with the relative merits of Baillargeon's thesis, and I have no reason to question her or others' experimental methods. Again, I am concerned with evaluating her *uses of words* as she interprets experimental results toward theorizing. One question is: How do we conceptualize the dependency of the abilities in terms of the supposed innate ideas, systems, or what have you? (Which words do we choose?) With these concerns in mind, it is worth establishing that my priorities stand in stark contrast to Baillargeon's strong emphasis on *experimental results* as giving the final word. She says: 'Although this point is often misunderstood by empiricist researchers, claims about innate ideas are of course empirical, and as such they are subject to revision in light of new experimental findings' (p. 2). The irony of this statement is palpable. There is no question that philosophical views frame the debate. (Otherwise, why summarize them at the start of her research review?) So, is Baillargeon saying that claims about innate ideas are empirical *only*? Maybe she is saying that they are at least partially empirical. Either way, are we to suppose that the acceptance, revision or rejection of those views must be based solely on experimental results? Well, they are not. To the extent that my investigation does speak to the issue of innate ideas, I will suggest why. But first we need to go over some details on the proposed underlying system.

14.2.1 The 'physical-reasoning system'

Baillargeon (2008) claims the young infant brain possesses a 'physical-reasoning system' or 'abstract computational system designed to monitor events as they unfold and to interpret and predict their outcomes…' (p. 3) When infants see an event, their physical-reasoning system 'builds a specialized representation of the event. Any information included in this representation is interpreted in terms of infants' core concepts and principles.'

The physical-reasoning system is innate or ready to go very soon after birth. So 'in the first weeks of life, an infant's physical representation of an event typically includes only basic information of the event' (p. 4). There are two forms of basic information: *spatiotemporal* (which 'specifies how many objects are involved in the event…and how their arrangement changes over time') and *identity* (which 'provides categorical or ontological information about each object, such as whether it is inert or self-propelled…and whether it is closed or open'). Experience brings

more information to physical representations of events, as infants 'identify the variables relevant for predicting outcomes. Variables are identified separately for each event category' (p. 4). Baillargeon claims that the earliest categories 'include *occlusion events* (object behind another object, or *occluder*), *containment events* (object inside container), *covering events* (object under cover), and *tube events* (object inside tube).' Event categorization requires interpretation. So, for example, in occlusion events a variable such as width 'calls infants' attention to the relative widths of objects and occluders and specifies that an object can be fully hidden behind an occluder if it is narrower, but not wider, than the occluder.' Thus, the occlusion event 'provides a rule' for how the infant is to interpret the event.

At this point, the researcher can organize variables into 'vectors' in the service of charting a developmental course of the abilities in question. So, physical-event 'representations' (or models) are charted as tree diagrams, with yes – no answers at each node. Baillargeon charts the vector 'When is an object that reappears behind an occluder the same object that disappeared?' At four months of age, the infant's physical reasoning system can determine whether an object that reappears is the same size as an object that disappeared. At seven and one-half months, it can determine whether the object that reappears has the same pattern as that which disappeared. A half month prior to the first birthday, colour comparisons are possible (see Wilcox, 1999).

Putting it all together, while watching an event, infants 'represent' its basic (spatiotemporal and/or identity) information. This information is used to categorize the event (e.g. as an occlusion event). Then infants

> tap their knowledge of the selected category, which lists the variables identified for the category. Information about these variables is then included in the physical representation and is interpreted in accordance with the variable rules and core knowledge. [So] ... while watching a red ball being alternately lowered behind and lifted above a screen, infants would first represent the basic information 'inert closed object being alternately lowered behind and lifted above inert closed object.' Infants would then categorize the event as an occlusion event, would access their knowledge of this event category, and would include information about all known relevant variables in their physical representation of the event. (Baillargeon, 2008, p. 4)

I suppose it is a sign of the times that we find such passages in scientific psychological journals. Anyway, it needs breaking down. Infants

'represent' an event's basic information. They use this information to categorize. They can categorize, have knowledge of categories and can select categories. Their categories list variables. Infants include variables in their physical representation and interpret the variables. Interpretations are carried out with respect to 'variable rules' and 'core knowledge.' (Again, Baillargeon claims that the fundamental bit of core knowledge is the 'principle of persistence.') Also, in the red ball example, infants represent 'inert', 'closed', 'object', 'lowered', 'behind', 'lifted', and 'above', and understand an event as involving 'occlusion.'

14.2.2 Parameters of conceptual investigation and objections from the research community

Anyone familiar with Wittgenstein's contributions to philosophical psychology will see that there is much in this manner of theorizing that begs conceptual investigation. Here, toward setting parameters for my own investigation, I want to make several observations about how Baillargeon uses words to describe the makeup and workings of the physical-reasoning system and then present a number of objections from others about the general tenor of her theorizing.

First, going back to the initial steps in Baillargeon's (2008) account, the workings of the physical-reasoning system are described via use of the psychological verbs 'to monitor', 'to interpret' and 'to predict.' She does not hesitate to describe the system in terms of psychological abilities. (After all, it is a physical-*reasoning* system.) Second, the system 'builds' specialized event representations – presumably put to use by the system – and *information* is included in the specialized representations. Third, the information is *interpreted* on the basis of 'core concepts and principles', without which the system would have nothing to go on. (This is why her theorizing is characterized as a 'principle-based approach.') Fourth, Baillargeon undoubtedly considers the workings of the physical-reasoning system to be automatic; that is, it is governed by material- and efficient-causal processes as the system interacts with environmental events. Yet, she uses words in such ways as to portray the system – and so infants – as being *actively engaged* in doing several things that require *skill* (e.g. using information, categorizing, interpreting). Finally, the infant and the system are portrayed as *being able to manipulate symbols in accordance with rules*. (Otherwise, how would infants be able to represent 'inert closed object' or anything else?)These observations form the backbone of what I have to say in the next section. But the purpose of my investigation – to identify sources of confusion – needs some justification in light of doubts and objections

about the general tenor of theorizing exemplified by Baillargeon's thesis, expressed already by members of the scientific community. A handful of examples, varying in emphasis, will do.

Being concerned with researchers' interpretations of infants' behaviours, Hood (2001) focuses on the pitfalls of attributing possession of knowledge states to infants 'in the absence of a commentary' (p. 1283). The point is similar to what we find in Coulter's (1983, pp. 108–14) more rigorous, philosophical discussion of 'opaque ascription contexts in behavioural research.' In such contexts, researchers who want to bridge the epistemological gap assign values to infants' behaviours by using concepts 'alien to the child and thus inaccurate to [the child's] way of seeing, perceiving or experiencing' (p. 109).[7] Another criticism emphasizes methodology and parsimony. Cohen and Marks (2002) question whether, for example, six- to nine-month-old infants possess some primitive counting ability (e.g. Starkey, Spelke and Gelman, 1983, 1990), including the ability to add and subtract (Wynn, 1992). They say researchers 'should be cautious about attributing sophisticated cognitive processes to young infants when simpler processes will suffice' and that 'caution and parsimony are the best principles to follow when trying to understand the development of infants' abilities' (p. 200).[8] Then there is the suggestion of alternative explanation. Focusing on research exploiting the 'violation of expectation' paradigm, Moore and Meltzoff (2008) challenge the assumption that behaviours exhibiting 'expectation' involve any sort of knowledge. They say looking-time experiments require 'complex chains of inference to bear on infants' understanding of object permanence' and prefer replacing 'knowledge' on the part of infants with 'perceptual expectations' (p. 169). Finally, as a criticism of note, Munakata et al. (1997) go right to the heart of theorizing by the likes of Baillargeon in suggesting that principle-based approaches to infant perception and cognition imply that 'knowledge takes the form of principles that function like propositions' and add that descriptions of behaviour conforming to principles are accepted 'as mental entities that are explicitly accessed and used in the production of behavior' (p. 687). In other words, principle-based theorists tend to describe infants' behaviour – say, in looking-time experiments – *in terms of the purported principles*, which at least implicitly take the form of mental entities 'accessed and used in the production of behavior.'[9] I suppose we can count this as a rejection of mentalism.

Although these and similar criticisms are worthy, I question their effectiveness because they do not cut to the sources of conceptual incoherence. I believe it is not enough to observe that Baillargeon and others

have gone beyond the information given, that they see the workings of the infant mind/brain through the eyes (and words) of the scientist,[10] that they should not try to bridge the epistemological gap and/or should be more cautious and parsimonious. I doubt that arguing for alternative, empirically-based explanations and eschewing mentalism will result in researchers watching their words. Most of these criticisms only characterize the nature of the alleged confusion. Maybe we need a shift to *identifying sources* of confusion.

14.3 Sources of confusion

14.3.1 Conflating natural abilities and acquired abilities

A Wittgenstein-informed developmental psychology observes the results of conceptual analyses of our uses of abilities and skills concepts that, rather than being aimed at informing or producing empirical theory, are aimed revealing sources of confused word-use.[11] Among other things, such analyses result in the distinction between natural powers and natural and acquired abilities. Also, the analyses suggest that some acquired abilities are based on natural abilities, exploited in implicit teaching and learning contexts. It is important to emphasize that the basis for the distinction is *grammatical*; that is, the distinction is expressed in the *rules of use* of our words pertaining to what some nonliving stuff can do (natural powers) and what some living beings can do (natural and acquired abilities).

What are the implications of how we talk about the various things that nonliving stuff and living beings can do?[12] When a boulder rolls down a hillside and topples a small tree, our description of this event does not reference teaching practices, learning and degrees of skill. What occurred exhibited the boulder's *natural powers*. Now consider the rooting and sucking reflexes of human infants. We call these *natural abilities* because they pertain to what some *living being* is able to do. No teaching and learning is required for display of these abilities. However, we can describe some natural abilities in terms of their foundational roles in the teaching and learning contexts that result in *acquired abilities*, associated with teaching practices, learning, degrees of skill and rule-following (or normative behaviour). Acquired powers are at the bottom of the possibility of any person-psychology.

Behavioural researchers exploit the natural abilities of rats and pigeons to shape behaviours in a Skinner box. We can say those behaviours are acquired, but not in the sense of acquired abilities possessed and exhibited by humans. On this point, consider Wittgenstein's (1953,

§244) description of a child learning 'new pain-behaviour' (or learning how to use words and conventional gestures to express their pain). This learning must work off natural abilities that are the expression of pain (e.g. crying, grimacing). There could be no way to get the teaching and learning of more refined, linguistic forms of pain-expression (e.g. 'My knee hurts!') off the ground without natural expressions of pain. (The same goes for the learning of some emotion concepts.) There are all sorts of teaching and learning contexts that work off of natural expressions and make possible, for example, the ability for persons to talk about the pain they had last week, to joke about their pain or hope they do not have the same pain again. The abilities and skills relevant to any thoroughgoing developmental psychology of human infants cannot be separated from the domain of normative (or conventional) behaviour.

The contexts in which natural abilities – in our example, as natural expressions of pain – are exploited in order for new, conventional expressions to be taught and learned are called 'primary language-games.' As a matter of grammar (or logic), participation in primary language-games is required for participation in 'secondary language-games' of pain-expression that are not based on natural expressions. A person simply cannot 'put on a brave face' while in pain 'without already having mastered the primary language-games that involve learning to use the word "pain" or related words' (Harré and Tissaw, 2005, p. 293). The distinction is relevant to the concerns of this chapter on at least two counts. First, it is relevant insofar as developmental researchers and theorists attribute to young infants – or physical systems within young infants' bodies – abilities and skills that cannot be acquired without such teaching and learning contexts. Second, it suggests why I emphasize the 'person' concept for the remainder of this section. Young human infants (and non-human animals) cannot engage in secondary language-games, because they have not acquired in primary language-games the requisite skills. For human infants, these skills will be acquired as part of language learning, which is part of becoming a person. (For non-human animals, no such status will be achieved.) Although I will mention these types of language-game only once hereafter, we should keep in mind their connections to natural abilities and acquired, conventional abilities (or powers).

In describing the workings of her physical-reasoning system, Baillargeon (2008) conflates innate, natural abilities and acquired abilities. She posits an innate system, whose workings she describes *as if the system were a person with acquired powers*. (Below I discuss this further as the second source of confusion.) The system interprets and predicts, uses

information, categorizes, and manipulates symbols according to rules. All of these are acquired – indeed *conventional* – abilities that can be carried out to varying degrees of skill. Yet, the system is supposed to be innate or in place very soon after birth. How can this be?

The primary source of Baillargeon's conflation of natural abilities and acquired abilities is deeply flawed logic. It makes no sense to ascribe to a part of a creature – whether that part is the brain or some system that is part of the brain – psychological abilities and skills that can be ascribed only to the creature as a whole. Bennett and Hacker (2003) have detailed how and why neuroscientists have committed this 'mereological fallacy.' Due to space constraints, I cannot elaborate the many ways in which their work applies to Baillargeon and many other researchers on infant cognitive development. It is enough to say that the researchers certainly commit the fallacy. Brains (or their parts) do not reason, interpret, predict, supply basic information or identify variables. Persons do these things.

As suggested in Section 14.1, some decades ago the floodgates were opened to researchers attributing remarkable cognitive abilities to very young infants. I also mentioned that in attempts to provide complete explanation of underlying mechanisms responsible for those abilities, Baillargeon and other researchers must wait for their experimental results to be integrated with – or validated by – evidence from the neurosciences. Perhaps Baillargeon's and other researchers' terminology is figurative; it is a terminology that waits for conversion to more precise terminology once the 'ultimate evidence' is in. Maybe Baillargeon does not want to ascribe conventional abilities to her physical-reasoning system, but *has to* – for now. Well, if it is right to consider her terminology to be figurative, then her theorizing stands apart from much theorizing in neuroscience. For the most part, neuroscientists ascribing psychological predicates to the brain have not and do not use the predicates in specialized ways (see Bennett and Hacker, 2003, p. 75). I do not think Baillargeon uses psychological predicates in specialized ways either. In any case, the truth is that she and other researchers *really do not have to attribute conventional cognitive abilities to a supposed innate system of the brain*. Baillargeon has made the choice to do so. But, then, her theorizing, too, is conventional; it is part of a decades-long trend in cognitive developmental science that needs continued, serious questioning.

14.3.2 Theorizing young human infants as persons

Human infants are not born persons. They *become* persons. *How* they become persons is a matter of empirical investigation. *That* they are not

born persons, but become persons, is shown in rules of use (grammar) of the concepts 'human being' and 'person.' We mix the concepts all the time in everyday contexts. 'You're a fine human being!' is not philosophically problematic. But it is confused to describe the workings of young infants' minds in terms of acquired powers possessed by persons. By extension, the same can be said of non-human animal cognition.

The category 'human being' is a 'substance concept' that specifies a category of animal with certain social, cognitive and physical characteristics and capacities; 'person', on the other hand, 'qualifies a substance concept of an animal of such-and-such kind, earmarking the individual of the relevant kind as possessing ... a distinctive range of powers, a personality and the status of moral being' (see Hacker, 2007, pp. 312–3). In other words, and limited to 'range of powers', reference to, and expression of, acquired abilities are part of what distinguishes 'person' from 'human being.' We can try to specify a boundary in terms of what abilities or skills must be acquired in order for a human infant to acquire personhood. But what would be the point? Everyone knows that in many ways, caregivers treat their infants as persons. Such treatment is fundamental to the training at personhood, and there is little or nothing empirically or philosophically problematic in conceptualizing human infants as 'developing persons' (see Tissaw, 2000, 2010, p. 166).

Sure, young infants *respond* to events, and it is true that their responses are based on the workings of an adequately functioning central nervous system. Certainly, it is within the purview of developmental science to reveal the biological 'systems' that make possible what infants at various points in their development can do. But Baillargeon's (2008) physical-reasoning system cannot do what she says it does; for neither the system itself nor the behaviour supposedly in evidence of the system (e.g. 'looking reliably longer', 'surprise') provide *criteria* adequate for the ways in which she theorizes young infants' cognitive abilities. As we have seen, she says young infants' physical-reasoning system is able to 'monitor', 'interpret', and 'predict.' In a sense, she might be right about monitoring. (We say, for example, that security cameras monitor.) But interpreting and predicting are species of *reasoning*. They are acquired cognitive abilities, exhibited by persons, that can be carried out to varying degrees of skill. Given any sort of experimental condition, no behavioural dependent variable exhibited by an infant is an adequate criterion for any kind of reasoning. That it would be simply is evidence of feeble theorizing which, by the way, contrasts with well-designed, even clever, experiments.

One problem is that in habituation and other forms of research on infant cognition, some researchers seem to assume that descriptions such as 'interested', 'disinterested', 'bored', 'noticed', and 'recognized' are meaningful in the same ways as they are used in describing cognitive activities, events and processes among adults. I suspect that researchers such as Baillargeon will see this as a naïve concern. In so many words, they might respond by saying, 'Of course, we do not assume that the words we use to describe infants' experiences and/or cognitive activities have the same meaning for the infant as they do for us. This is why we say that infants operate in accordance with core concepts or principles!' But the point I'm driving at is subtler. Imagine a teacher instructing a pupil on the meanings of words such as 'interest', 'recognizing', 'boredom', and 'knowing' by pointing to examples of people who are interested in something, recognizing something, are bored and know something. Whatever the teacher points to, they cannot be examples of experiences, states or processes. We cannot learn the meanings of these words via ostension alone.[13] The meanings of the words, shown in varieties of their use, are learned in activities carried out by persons with the requisite linguistic skills. Overt behaviour that looks like 'surprise' does not equal surprise *to the infant as an experience*. Yet, again and again, researchers such as Baillargeon use psychological concepts to make attributions of abilities, as if young infants have achieved personhood.

At this point, it may be instructive to consider the very difficult question of what it is to 'conceptualize', 'possess a concept' or 'have a concept of.' One of the lessons I have taken from Wittgenstein is to avoid beginning an inquiry of this sort in hopes of answering a question such as 'What is a concept?' Rather, the inquiry should aim *to investigate what non-concept users cannot do and what concept users can do*. On this question, I defer to three observations made by Hacker (2007). First, the 'horizon' of the thinking of non-human animals 'is determined by the limits of their [overt] behavioural repertoire. They can intelligibly (truly or falsely) be said to think only what they *can* express in their behavior' (p. 237). Non-human animals and young infants can be said to think, but not in the sense of thinking by employing concepts. (Thus, for example, it is nonsense to say that non-human animals and human infants can count.) Second, 'If a creature has mastered a language with logical connectives and quantifiers, then and only then is it possible for it to conceive of general truths, to think both of how things are and how they are *not*, to think both of what exists and of what does not exist, to think conditional thoughts, and with the aid of tensed verbs and modal expressions, counterfactual thoughts' (p. 238). Third, we should resist

taking the remarkable recognitional abilities and discriminatory powers of many animals – including young human infants – as justification for ascription of concept use (or possession) to them. For such abilities and powers are

> not sufficient for possession of any concept, and not even necessary for some concepts. It is necessary for possession of the concept of red. But a being to whom possession of this concept may be ascribed must grasp that red is a colour; that it applies to extended objects, but not to sounds or smells; that if something is red all over, it cannot simultaneously be green all over; that red is darker than pink, and more similar to orange than yellow, and so on. In short, [s/he] must grasp the logical articulations of the concept, i.e. the rule-governed use of a word that expresses it. To be sure, concept possession is a complex ability, which admits of degrees. Is mere recognitional ability sufficient to ascribe a minimal master of the concept? That is a matter of decision, but there is little to be said for decided in favour of the proposal. It would detach the concept of concept possession from the cluster of abilities with which is normally and usefully associated, and detach the concept of a concept from its connection with the concepts of application and misapplication, use, misuse and abuse, subsumption, extension, replacement and substitution. (Hacker, 2007, fn., pp. 239–40)

It is important to reiterate that the philosophical method that informs this passage does not aim toward informing or producing empirical theory. What Hacker says about 'concept possession' is the result of conceptual analysis on the *possibilities* of concept use (or 'conceptualization'). It is, as he suggests, a *choice* to attribute, for example, 'attenuated' conceptual abilities to young human infants or non-human animals. But the results of his analyses suggest why this might be a bad idea. Should we theorize on attenuated conceptual abilities when we have yet to understand what it means to be a full-fledged concept-user?

The immediately foregoing brings to mind my previous observation that Baillargeon and other researchers simply have made the choice to attribute conventional cognitive abilities to young infants, their brains or 'representational' systems. To say again, 'looking reliably longer' or exhibiting facial expressions of 'surprise' cannot count as criteria for infants 'representing' any sort of basic information such as 'inert closed object being alternately lowered behind and lifted above inert closed object' (Baillargeon, 2008, p. 4). It makes no more sense to say this of

a physical-reasoning system than it does of a young infant. What is to keep us from saying that an earthworm is wriggling on a hot sidewalk because its physical-reasoning system 'knows that if it does not get off the sidewalk, it will dry up'? Now contrast this with a barefoot person who can tell us why they choose to stand on the grass.

14.4 Conclusions

It is a worthy cause to attempt to explain young infants' responses to variable manipulations in habituation and other forms of experiments by theorizing the physical foundations of the responses. But the cause and the methods that serve it do not license attribution of the abilities and skills that conceptually mark the domain of personhood to young infants or the physical foundations. In explaining results of habituation and other forms of research on young infants' cognitive development, researchers should not choose words that conflate natural and acquired abilities and thus portray infants as possessing abilities and skills that set apart the category 'person' from 'human being.'

Part of what licenses continuing attribution of personhood to young infants is a decades-long tradition of making such attributions – not necessarily directly, but indirectly describing the physical foundations in terms of psychological predicates. The claims of innate or core 'principles' are just an extension of the licensure, sometimes, I suspect, in service of misguided scientism. Whether or not the scientists want to acknowledge it, philosophy will always have a place in *good* scientific psychological theorizing. For whatever the methods employed, the research alone cannot ensure prudence in selecting the right words.

In describing young infants' responses to variable manipulations in the laboratory, researchers must always consider the possibility that the words they choose to describe the responses probably are not meaningful in the same senses as when they are applied to persons. When applied to a young infant in the laboratory, 'surprise' does not carry the potential meanings (or uses) as when applied to or expressed by persons. Potential for conceptual incoherence lurks in a simple, circular justification that connects 'surprise' with the dependent variable looking-time. For example: 'Because infants' surprise at an event typically manifests itself by prolonged attention to the event, it is assumed that, if infants are surprised by the impossible event, they will look reliably longer at it than at the possible event' (Baillargeon, 1994, p. 9). What researchers call 'surprise' on the part of infants in laboratory contexts seems to me a natural, expressive response, devoid of psychological

'content.' It is a response that later will be exploited by adults in the teaching and learning contexts of primary language-games. In general, surprise is marked by involuntary response (Wittgenstein, 1953, §628; see also Hacker, 1996, p. 611). This is why, given various contexts, we say that non-human animals, infants, and adults are (or were) surprised. All sorts of factors inform these attributions. Otherwise, we could not make them. And when we admit of our own surprise, we may cite some of the factors in secondary language-games. But young infants cannot and do not do this. Nor can they show us evidence of associating their surprise with their expectations, assumptions, past experiences or the role of antecedent events.

Which words should researchers like Reneé Baillargeon choose? As I see it, there are two options. 'Looking-time' is a good dependent variable. For lack of a better way of putting it, it is 'psychologically neutral' or carries limited or no psychological baggage. Or, if researchers must use words like 'surprise', they should *always* qualify their use of the concept in the writing of research reports. Among other things, this means they should acknowledge that they know it makes no sense just to be surprised. One must be surprised *at something*. And no young infant possesses the acquired abilities needed to conceptualize that they have been surprised, for example, when a tall cylinder became almost fully hidden when positioned behind a short container (cf. Hespos and Baillargeon, 2001). I believe four-and-a-half-month-old infants do not reason about occlusion events in this way. In fact, I believe there are no conceptual or scientific bases to conclude they reason at all.

Notes

1. In their review of visual habituation in the study of infant cognition and learning, Colombo and Mitchell (2009, p. 226) claim to have turned up 777 publications between 1962 and 2008 in a PsychInfo search using 'habituation' as a keyword with 'prenatal,' 'neonatal,' 'infants', and 'preschoolers' as population parameters.
2. Discussed in Santrock (2011, p. 156).
3. One paradigmatic example is the orienting reflex in Sokolov's (e.g. 1966) 'comparator model' of habituation.
4. For example, Berkson and Fitz-Gerald's (1963) research on infant chimpanzees.
5. At least as it is understood by Baillargeon, this attitude toward metaphysics is in keeping with the overall tenor of the special issue of *Perspectives on Psychological Science* in which Baillargeon's article appears, entitled 'From Philosophical Thinking to Psychological Empiricism.'

6. For example, Wilcox (1999) concludes an account of four experiments on young infants' shape, size, pattern and colour feature individuation with a summary of 'evidence from the neurosciences' which, combined with her findings, 'suggest neural mechanisms for early processing biases' (p. 162).
7. Coulter's point about opaque ascription contexts calls to mind James's (1890/1983, p. 195) 'psychologist's fallacy'. As I have discussed elsewhere (Tissaw, 2007, pp. 232–3), researchers risk generating conceptual incoherence when they explicitly or implicitly assign to infants and young children cognitive states, processes or abilities that they cannot or do not conceptualize themselves because they lack the ability to use language.
8. Cohen (2002) adds that the infants in Elizabeth Spelke's 'number experiments' responded only to changes that violate expectations and that such responses were not based on counting procedures. See Geary (2006) for a discussion of theoretical debate on the arithmetical competencies of infants.
9. Munakata et al. (1997) offer the following analogy: '…[O]ne could say that infants' behaviour in a looking-time task accords with a principle of object permanence, in the same way that one could say that the motions of planets accord with Kepler's laws.' To 'conclude that infants actually access & reason with an explicit representation of the principle itself' would be like explaining 'the motions of the planets by claiming that the planets derive their next location in space on the basis of reasoning with Kepler's laws.'
10. This tendency is evident, for example, in Baillargeon (2001) attributing to infants the ability to 'identify a variable' while under some experimental conditions they would 'have difficulty collecting data' about objects (p. 349) or, in other conditions, 'more easily collect qualitative data' (p. 350).
11. Providing an explanation of the whys, wherefores and hows of Wittgenstein's philosophical method and its relevance to empirical psychological research and theorizing is beyond the scope of this chapter. Harré and Tissaw (2005, ch. 7) provide an account that is tuned to psychological theorizing.
12. The remainder of this paragraph summarizes an account by Harré and Tissaw (2005, pp. 84–5).
13. See Harré and Tissaw (2005, pp. 79–80).

References

R. Baillargeon (1994) 'Physical Reasoning in Young Infants: Seeking Explanations for Impossible Events', *British Journal of Developmental Psychology*, 12, 9–33.

—— (2001) 'Infants' Physical Knowledge: Of Acquired Expectations and Core Principles' in E. Dupoux (ed.) *Language, Brain, and Cognitive Development: Essays in Honour of Jacques Mehler* (Cambridge, Mass.: The MIT Press), 341–61.

—— (2008) 'Innate Ideas Revisited: For a Principle of Persistence in Infants' Physical Reasoning', *Perspectives on Psychological Science*, 3, 2–13.

M. R. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).

G. Berkson and F. L. Fitz-Gerald (1963) 'Eye Fixation Aspect of Attention to Visual Stimuli in Infant Chimpanzees', *Science*, 139, 586–7.

N. Chomsky (1965) *Aspects of the Theory of Syntax* (Cambridge, Mass.: MIT Press).

L. B. Cohen (1966) 'Observing Responses, Visual Preferences and Habituation to Visual Stimuli in Infants', *Dissertation Abstracts*, 27, 310.
—— (2002) 'Extraordinary Claims Require Extraordinary Controls', *Developmental Science*, 5, 211–2.
L. B. Cohen and K. S. Marks (2002) 'How Infants Process Addition and Subtraction Events', *Developmental Science*, 5, 186–212.
J. Colombo and D. Mitchell (2009) 'Infant Visual Habituation', *Neurobiology of Learning and Memory*, 92, 225–34.
J. Coulter (1983) *Rethinking Cognitive Theory* (New York: St. Martin's Press).
R. L. Fantz (1964) 'Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones', *Science*, 146, 668–70.
D. C. Geary (2006) 'Development of Mathematical Understanding' in D. Kuhl and R. S. Siegler (eds) *Cognition, Perception, and Language*, vol. 2 in W. Damon (series ed.) *Handbook of Child Psychology*, 6th edn (New York: John Wiley and Sons), 777–810.
P. M. S. Hacker (1996) *Wittgenstein: Mind and Will. An Analytical Commentary on the Philosophical Investigations*, vol. 4 (Oxford, UK: Blackwell).
—— (2007) *Human Nature: The Categorial Framework* (Oxford, UK: Blackwell).
R. Harré and M. A. Tissaw (2005) *Wittgenstein and Psychology: A Practical Guide* (Aldershot, UK: Ashgate).
S. J. Hespos and R. Baillargeon (2001) 'Infants' Knowledge About Occlusion and Containment Events: A Surprising Discrepancy', *Psychological Science*, 12, 140–7.
B. Hood (2001) 'Guest Editorial: When Do Infants Know About Objects?' *Perception*, 30, 1281–4.
W. James (1983) *The Principles of Psychology* (Original work published 1890) (Cambridge, Mass.: Harvard University Press)
L. A. Kahn-D'Angelo (1987) 'Infant Habituation: A Review of the Literature', *Physical & Occupational Therapy in Pediatrics*, 7, 41–55.
M. K. Moore and A. N. Meltzoff (2008) 'Factors Affecting Infants' Manual Search for Occluded Objects and the Genesis of Object Permanence', *Infant Behavior & Development*, 31, 168–80.
Y. Munakata, J. L. McClelland, M. H. Johnson, and R. S. Siegler (1997) 'Rethinking Infant Knowledge: Toward an Adaptive Process Account of Successes and Failures in Object Permanence Tasks', *Psychological Review*, 104, 686–713.
J. W. Santrock (2011) *Child Development*, 13th edn (New York: McGraw-Hill).
A. Slater, V. Morison, and M. Somers (1988) 'Orientation Discrimination and Cortical Function in the Human Newborn', *Perception*, 17, 597–602.
E. N. Sokolov (1966) 'Orienting Reflex as Information Regulator' in A. Leontyev, A. Luria, and A. Smirnov (eds) *Psychological Research in the U.S.S.R.* (Moscow: Progress Publishers), 334–60.
E. S. Spelke (1988) 'Where Perceiving Ends and Thinking Begins: The Apprehension of Objects in Infancy' in A. Yonas (ed.) *Perceptual Development in Infancy* (Hillsdale, N.J.: Lawrence Erlbaum Associates), 197–234.
P. Starkey, E. S. Spelke, and R. Gelman (1983) 'Detection of Intermodal Numerical Correspondences by Human Infants', *Science*, 222, 179–81.
—— (1990) 'Numerical Abstraction by Human Infants', *Cognition*, 36, 97–127.
M. A. Tissaw (2000) 'Psychological Symbiosis: Personalistic and Constructionist Considerations', *Theory & Psychology*, 10, 847–76.

—— (2007) 'Making Sense of Neonatal Imitation', *Theory & Psychology*, 17, 217–42.
—— (2010) 'A Critical Look at Critical (Neo)Personalism: "Unitas multiplex" and the "Person" Concept', *New Ideas in Psychology*, 28, 159–67.
T. Wilcox (1999) 'Object Individuation: Infants' Use of Shape, Size, Pattern, and Colour', *Cognition*, 72, 125–66.
L. Wittgenstein (1953) *Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Blackwell).
K. Wynn (1992) 'Addition and Subtraction by Human Infants', *Nature*, 358, 749–50.

15

The Unconscious Theory in Modern Cognitivism

Alan Costall

Many of the conceptual confusions fundamental to modern cognitivist theory had already been identified and widely recognized *before* the 'cognitive revolution' of the 1960s. Yet, whenever such confusions are pointed out, they are either fleetingly acknowledged, only to be quickly forgotten, or, more usually, emphatically denied. And, as I have found to my own cost, cognitive psychologists become outraged if you suggest that they may even be *dualists*.

The trouble is that when psychologists think of *theory*, they usually assume that, in the case of theory, surely everything is set out explicitly. Everything is *up-front*. Yet, as I will try to explain, some fundamental aspects of modern cognitive theory are *implicit* in (1) metaphors, (2) methodology, (3) long-established scientific terminology, and even (4) the official history of modern cognitivism. They are *unconscious*. So, when cognitive psychologists flatly *deny* that they are confused, they are, in fact, *in denial*. They are saying one thing, and meaning quite another.

My chapter, then, is concerned with the precarious conceptual structure of modern cognitivism. But it is not quite conceptual analysis in the sense in which this has developed within analytical philosophy. It is an attempt to situate the language of psychological theory and methodology historically within research practice. It draws upon discursive approaches to the history of psychology, most notably the work of Kurt Danziger (1990). But it is also, I believe, consistent with Wittgenstein's insistence that language use needs to be situated within 'the broader "forms of life" which ultimately give those language-games their significance' (Toulmin, 1969, p. 61). I will focus upon three examples of unconscious theory that continue to structure modern cognitivist theory.

15.1 The computer metaphor, the 'active mind', and mind-body dualism

My first example of unconscious theory concerns the unintended and unanticipated implications of metaphor. Since the earliest days of cognitive psychology, a contrast has been drawn between the poverty of the mechanistic and associationist approach of the neo-behaviourists, and the alternative emphasis upon the 'active mind'. According to this view, perceiving, remembering and thinking should be understood, as Bartlett (1932) nicely put it, as 'an effort after meaning'.

One of the main attractions of the computer metaphor of mind has been that it promises to capture the active and organized nature of psychological processes. As its early proponents themselves insisted, the so-called *computer* metaphor, as it has been explicitly formulated in modern psychology, should be regarded as a *program* metaphor:

> Our position is that the appropriate way to describe a piece of problem-solving behaviour is in terms of a program, a specification of what the organism will do under varying environmental circumstances in terms of elementary information processes it is capable of performing. *This assertion has nothing to do – directly – with computers.* Such programs could be written (now that we have discovered how to do it) if computers never existed. ... Digital computers come into the picture only because they can, by appropriate programming, be induced to execute the same sequences of information processes that humans execute when they are solving problems. (Newell, Shaw, and Simon, 1958/1991, p. 388, emphasis added)

In their early work on the 'Logic Theorist', Newell and Simon used their relatives and graduate students to 'run' their program by following the instructions contained in the program they were given (see Gigerenzer and Goldstein, 1996). And here is the crucial 'twist': The *people* running the program did not themselves know what they were doing: 'The actors were no more responsible than the slave boy in Plato's *Meno* [who "solved" a geometrical problem despite himself], but they were successful in proving the theorems given them' (Simon, 1991, p. 207).

The implications of the program metaphor have been endlessly discussed, and so I will not to pursue all the different issues here, but simply two of the more neglected ones. The first is that the program metaphor actually *undermines* the cognitive theorists' attempt to theorize the active mind, the very reason they invoked this metaphor. As

Sigmund Koch pointed out long ago, the computer metaphor of mind presents us with a 'thoroughly rule-regulated' image of cognition that it can find no place for the 'cognizer' (Koch, 1981, p. 258).

Thanks to the reassurance of the program metaphor, cognitive psychologists simply took the *primacy* of rules and representations for granted (see Dreyfus, 1995). So, from then on, their *basic* question became not *how* rules and representations come to play their role in human life in the first place, but *which* particular rules and representations might explain certain aspects of psychological functioning.

In this way, they begged two fundamental questions. The first, as I have already indicated, concerns *the conditions of possibility* of representation. The second is that the effective use of rules and representations depends upon their intelligent deployment – upon intelligence that is not, itself, contained in the rules, but precedes and transcends them. In real life, as opposed to the abstract and simplified settings envisaged by the cognitive theorists, the *blind* following of rules would hardly constitute intelligence, but mindless stupidity (Shaw, 2003).

The second neglected implication of the computer metaphor concerns an implicit *dualism* of body and mind quite at odds with its celebrated role as the antidote to such a dualism where 'brain and mind are *bound* together as computer and program', or hardware and software (Johnson-Laird, 1988, p. 23, emphasis added). This is why Pylyshyn could claim that cognitivism provides us with 'a science of structure and function *divorced* from material substance' (Pylyshyn, 1986, p. 68, emphasis added).

I first raised this issue some years ago, but to little effect (Costall, 1991). The problem is that psychologists, and even critics of the computer metaphor, have been so focused upon the software or program aspect of the computer metaphor, that they hardly bothered to consider what precisely the *computer* or *hardware* was supposed to represent, not least, whether it refers to the mind, the brain or the body. Admittedly, some theorists have invoked aspects of the hardware as part of the computer metaphor, such as the central processing unit, memory stores, and buffers. But this was not what the proponents of the *program* metaphor had in mind: 'we are *not* comparing computer structures with brains, nor electrical relays with synapses' (Newell, Shaw, and Simon, 1958/1991, p. 388, emphasis added). What they had in mind, instead, was the *ideal* of a computer as a 'general purpose machine' that formally underpins the separability of software from any specific hardware. This is a machine whose function is *entirely* unconstrained by the hardware. According to this ideal, therefore, the hardware – the *body* – has no explanatory relevance at all

any more than does the paper on which physicists write their differential equations (see Costall, 1991, 2007). And so, despite the initial good intentions, the computer metaphor has come to embody mind-body dualism, and thereby *disembody* the mind.

15.2 Stimulus-response dualism

My second example of unconscious theory concerns how cognitive psychologists perpetuate something they emphatically claim to have undermined. The cognitive revolution is supposed to have entirely eliminated mechanistic, stimulus-response behaviourism.

One of the ways that John Broadus Watson (1913) attempted to sell behaviourism to psychologists was to promise to rewrite psychology in the language of stimulus and response. In fact, modern stimulus-response psychology was largely a Russian invention, and its ambitions were remarkable. As Vygotsky wryly observed, 'Everything – sleep, thought, work, and creativity – turns out to be a reflex' (Vygotsky, 1926–1927, cited in Packer, 2008, p. 12). This was hardly a spoof. For example, the work of Clark Hull, an influential American behaviourist, was mainly based upon rats running through mazes, and yet he began his influential book, *Principles of Behaviour*, with the following solemn, yet frankly crazy, claim:

> As suggested by the title, this book attempts to present in an objective, systematic manner the primary, or fundamental, molar principles of behaviour. It has been written on the assumption that all behaviour, individual and social, moral and immoral, normal and psychopathic, is generated from the same primary laws; that the differences in the objective manifestations are due to the differing conditions under which habits are set up and function. *Consequently the present work may be regarded as a general introduction to the theory of all the behavioural (social) sciences*. (Hull, 1943, p. v, emphasis added)

Since the rise of cognitivism, psychologists have claimed that they have *eliminated* stimulus-response thinking from psychological theory and practice. I only wish this were true. Their accounts of both human development and evolution presuppose a purely stimulus-response organism as their starting point – a 'primitive' organism that only eventually comes to psychological life through the intervention of 'cognitive processes'. But they also reinstate stimulus-response thinking in their

statement of the very task of cognitive theory, namely, to explain the mediating cognitive processes that occur *between* the stimulus and response:

> Partly, [the cognitive revival] resulted from *a recognition of the complex processes that mediate between the classical 'stimuli' and 'responses'* out of which stimulus-response theories hoped to fashion a psychology that would by-pass anything smacking of the 'mental'. The impeccable peripheralism of such theories could not last long. (Bruner, Goodnow, and Austin, 1956, p. vii, emphasis added)
>
> It is a basic premise of the 'cognitive revolution' in psychology that individuals do not respond directly to stimuli from the external environment but to their perceptions and cognitive interpretations of those *stimuli*. (Brewer and Hewstone, 2004, p. xi, emphasis added)

This, of course, is no alternative to the stimulus-response paradigm, but a variant of that very paradigm. And it is not as though this mediational version of the stimulus-response formula is new. The neo-behaviourists had already seen their task as that of inferring internal processes that might mediate between the stimulus and response. Its first influential statement was Woodworth's (1918) *Dynamic psychology*.

Remarkably, one of the most insightful criticisms of stimulus-response thinking was published *before* this schema had taken its hold – first consciously and then, later, *unconsciously* – on psychological theory. This was John Dewey's famous article on the reflex-arc concept. Interestingly, Dewey was already addressing a *meditational* version of the stimulus-response formula:

> the older dualism of body and soul finds a distinct echo in the current dualism of stimulus and response. Instead of interpreting the character of sensation, idea and action from their place and function in the sensory-motor circuit, we still incline to interpret the latter from our preconceived and preformulated ideas of *rigid distinctions between sensations, thoughts and acts*.
>
> We ought to be able to see that the ordinary conception of the reflex arc theory, instead of being a case of plain science, is a survival of the metaphysical dualism, first formulated by Plato, according to which the sensation is an ambiguous dweller on the border land of soul and body, the idea (or central process) is purely psychical, and the act (or movement) purely physical. (Dewey, 1896, pp. 357–8, 365, emphasis added)

Donald Hebb, in his presidential address to the American Psychological Society, has been one of the few psychologists over the years to be clear about the fundamental dependence of cognitive theory upon the stimulus-response paradigm:

> the whole meaning of the term 'cognitive' depends on [the stimulus-response formula], though cognitive psychologists seem unaware of the fact. The term is not a good one, but it does have meaning as a reference to features of behaviour that do not fit the S-R formula; and no other meaning at all as far as one can discover. The formula, then, has two values: first, it provides a reasonable explanation of much reflexive human behaviour, not to mention the behaviour of lower animals; and secondly, it provides a fundamental analytical tool, by which to distinguish between lower (noncognitive) and higher (cognitive) forms of behaviour. (Hebb, 1966, pp. 736–7)

Most psychologists are simply unconscious of the fact that they are committed to a variant of stimulus-response theory. This commitment is implicit in the terminology of 'stimulus and response', of 'input and output' and also of 'independent and dependent variables'. But linguistic revision would not be enough. *Stimulus-response theory is embodied in the standard experimental paradigms where 'conditions' are imposed upon 'subjects'*. The task of the subjects in such experiments is to *react* to the conditions imposed upon them, and emphatically *not* to choose their own conditions or transform them. During the experiment, the subjects, in effect, lend their minds and agency to the experimenter. In real life, as Dewey had argued, the so-called 'stimuli' do not normally *precede* the so-called 'responses' but come into play in the very course of our activity, and only function as 'stimuli' in relation to that activity.

15.3 Methodological behaviourism as mind-behaviour dualism

So far, I have considered ways in which psychological terminology, metaphor and experimental method have structured psychology theory not only in unconscious ways, but often in ways that also completely undermine the stated intentions of the theorists. I now want to consider an extreme example of unconscious theory. This concerns an *historical myth* about the origins of modern scientific psychology that has structured modern psychological thought in the most fundamental way possible.

This myth embodies an unrecognized *theory* of how psychologists themselves *do* psychology.

J. B. Watson established the main substance of this myth in his attempt to promote both himself and behaviourism. Watson's version consists of the following claims (see Costall, 2006):

(1) At the beginning scientific psychology had been dominated by introspectionism, a psychology based exclusively on self-reports (on the dualistic assumption that only these could provide proper evidence about mind);
(2) Introspectionism had become conclusively discredited as a scientific method;
(3) Watson and his fellow behaviourists were going, *in their actual research practice*, to replace introspectionism as the study of the mind, with methodological behaviourism, the study of 'mere' meaningless movements.

As I have argued elsewhere (Costall, 2006), *all* of these historical 'facts' are inventions.

(1) Early scientific psychology was *not* exclusively dependent on introspectionism. If anything, it was the era of 'brass-instruments' psychology. Watson's critics were well aware that he was exaggerating. Here is Thorndike complaining about the misleading central message of Watson's textbook, *Behaviour: An introduction to Comparative Psychology* (1914):

> The student is likely, as Watson's book stands, to be left with the impression that mental chemistry – the analysis of conscious states into elements ... – has been the regular, orthodox thing in human psychology. On the contrary objective methods and results have characterized a very large proportion of the work of recognized psychologists for thirty years. Ebbinghaus' *Memory* and Cattell's studies of reaction-time, for example, are as behaviouristic or objective as Bassett's study of rats or Yerkes' study of frogs. (Thorndike, 1915, p. 463; see also Woodworth, 1931, p. 62)

Early scientific psychology was *not* dominated by introspectionism, nor were most of the proponents of the method of introspection committed to a dualistic conception of this method. They did not regard it as the sole means of studying mind. After Freud, how could they?

(2) Introspection was *not* found wanting, nor was it abandoned following Watson's intervention. Introspection has continued to play a

crucial role in both scientific and clinical psychology, if under the less provocative name of 'self report' (see Kroker, 2003).

A fundamental dispute did take place in the early part of the twentieth century between two major schools of introspection (Würzburg and Cornell) concerning the role of imagery in thinking. This dispute is widely represented in the textbooks as evidence of the inherent unreliability of the method of introspection. However, *if* introspectionism really had been the hopelessly idiosyncratic method it is still widely claimed to be, how is it there could have developed collective *schools* of introspectionism, whose members could agree – at least among themselves – about their own methods and findings (Brock, 1991)? In fact, Monson and Hurlburt (1993) studied the research records of both laboratories and found that the findings themselves were largely in agreement. The fundamental dispute was not, after all, about the method, but about the *interpretation* of the results.

(3) Despite their rhetoric, Watson and his followers were *not*, in their actual research, methodological behaviourists. First of all, Watson remained in two minds about introspection, even though he tried to re-brand it as 'self-observation', 'word response', or 'thinking aloud'. Here is a remarkable passage from Watson's book, *Behaviourism* (1924):

> In many spheres of psychology and especially in psychiatry, self-observation, which is usually expressed in words by the subject, is the only kind of observation at our immediate disposal. The patient comes to the psychiatrist and says: 'I feel "sad" and "gloomy"'; or, 'Doctor, I am under a terrible strain – I fear I am going to kill my wife and children.' ... The physician then by a series of skilful questions begins to take the word responses of the patient. These responses, however, are from the physician's standpoint as objective as would be a moving-picture photograph of the subject's activity in weaving a rug or basket. (Watson, 1924, p. 42)

Watson's early critics certainly had fun pointing out the unacknowledged basis of Watson's confident claim that thinking consisted of nothing but kinaesthetic sensations (e.g. Watson, 1920):

> It is worth noting that introspection must have led Mr Watson to the conclusion that his mental imagery consists in kinaesthetic sensations mainly from his vocal apparatus. We are not justified in saying that it was bad introspection on his part: some minds may indeed be so poorly furnished with certain elements of enjoyment that their

> possessors live in a world divested of the glow of inner colour and the harmony of inner sounds. But the more fortunately endowed will reproach them for making their individual limitations the universal law. (Washburn, 1915, p. 212)

Karl Lashley, who had worked with Watson, divulged the following subversive secret to his friend, Edmund Jacobson (the pioneer of the clinical technique of progressive relaxation, a technique dependent on introspection): 'Lashley told me with a chuckle that when he and Watson would spend an evening together, working out principles of behaviourism, much of the time would be devoted to introspection' (Jacobson, 1973, p. 14).

In their influential text, *Protocol Analysis: Verbal Reports as Data,* Ericsson and Simon (1984) credited Watson with introducing the 'think aloud' method that they had used in their own work. But Herbert Simon was not drawing a contrast between Watson's method and introspection. In an interview with Baars (1986, pp. 365–6), he also acknowledged the close relationship between his own self-report studies and Otto Selz's early introspective work at Würzburg (See van Strien and Faas, 2004, for more on Selz).

I have dwelt upon Watson's *actual* dealings with introspection because they are not widely known, and yet undermine his *rhetorical* rejection of the method of introspection. This rhetorical rejection is widely represented in the modern textbooks as an established truth, and, despite the fact that self-observation is still widely used within psychology, including experimental psychology, a lot of nonsense continues to be written about introspection as a hopelessly unscientific, *self-enclosed* method. Watson's misrepresentation of introspection as a dualistic method is still highly influential, and underpins, in a deeply paradoxical way, modern psychology's unconscious commitment to mind-behaviour dualism.

Let us now turn to Watson's equally problematic *Behaviourism*. Watson's contemporary critics were well aware that the version of behaviourism he was promoting – in his rhetoric if not in his practice – was not an alternative to traditional mind-body dualism, but the other side of the same coin:

> Embedded in the very core of the behaviourist's doctrine is the Platonic distinction between mind and matter; and behaviourism, like Plato, regards the one term as real and the other as illusory. Its very case against dualism is stated in terms of that distinction and is made by the classical metaphysical procedure of reducing the one

> term to the other. This metaphysical distinction, rather than empirical evidence, is the basis on which behaviourism accepts or rejects data for scientific consideration and on which it forms conceptions for dealing with them. (Heidbreder, 1933, pp. 267–8; see also Dewey, 1914/1977, p. 445)

In fact, Watson was deeply inconsistent in his own unconscious commitment to mind-behaviour dualism. As Tolman (1932, p. 6) pointed out it, Watson 'dallied with two different notions of behaviour, though he himself has not clearly seen how different they are'. The first notion refers to mindless, meaningless physical movements; the second to mindful, meaningful action. *Officially*, the behaviourists were supposed to restrict themselves to studying behaviour in the former, mechanistic sense. *In practice*, however, what they were investigating were already psychologically meaningful activities, rats trying to find their way through mazes, cats working out how to escape from puzzle boxes.

Even the definitions of what kinds of behaviour belonged to the domain of behaviourism were remarkably inclusive, and, as we shall see, sometimes even quirky. (See Kitchener, 1977, for an extensive review of the different and often ambiguous meanings of 'behaviour' in the behaviourist literature.) Here is one of Watson's own attempts:

> By response we mean anything the organism does – such as turning toward or away from the light, jumping at a sound, and more highly organized activities such as building a skyscraper, drawing plans, having babies, writing books, and the like. (Watson, 1925, p. 7; cf. Tolman, 1932, p. 8)

I have already discussed the compromise that modern cognitive theory made with stimulus-response behaviourism, a compromise that is consistently denied. However, the cognitive psychologists also made a much more fundamental compromise with Watsonian behaviourism, one that has been openly celebrated in the textbooks for many years, as a grand scientific *synthesis*.

Despite the talk about 'cognitive revolution', cognitive psychologists uncritically accepted Watson's two-stage history of introspectionism giving way to reductionistic behaviourism, and simply added their own culminating stage. Psychology has again become the study of mind, but, this time, it is emphatically *not* based upon introspection, but, instead, upon the rigorous methodology developed by the behaviourists.

One of the earliest accounts of this synthesis comes from the second edition of Donald Hebb's psychology textbook:

> If Watson's work is seen as [a] house-cleaning operation..., its importance becomes clearer. ... In 1913 the whole case for mental processes seemed to depend on introspection; if it did, the case was a bad one, and 'mind' had to be discarded from scientific consideration until better evidence could be found ... *Paradoxically, it was the denial of mental processes that put our knowledge of them on a firm foundation, and from this approach we have learned much more about the mind than was known when it was taken for granted more or less uncritically*. (Hebb, 1966, pp. 5–6, emphasis added)

This narrative of synthesis continues to appear in the textbooks:

> Although cognitive psychologists now study mental structures and operations, they have not gone back to the introspective methods that structuralists such as Wundt employed. *They use objective research methods, just as behaviourists do*. (Carlson et al., 2006, p. 91, emphasis added)

I hope I have said enough about the behaviourists to explain why this claimed *synthesis* is a myth. The simple reason is that the components of the synthesis are a myth. The behaviourists were never, in practice, methodological behaviourists, and the cognitive psychologists have themselves no more been observing 'mere' behaviour than the behaviourists before them. Even when the participants in their experiments are 'merely' pressing buttons or computer keys (which is often the case), the experimenter has already negotiated the *significance* of the button-pushing prior to the onset of the experiment: 'press this key if this display looks brighter than this other one, and this key if it looks dimmer.'

So, why did the early cognitive psychologists, given their talk of scientific revolution, so readily accept Watson's mythical history and incorporate it into their own three-stage origin myth? The main problem is the conservatism of the existing methodology – even if, in the case of methodological behaviourism, it is a *mythical* methodology.

It is one thing to develop a novel approach; it is another thing, on the basis of that approach, to obtain a PhD, or an academic post, or publications and research grants. Such crucial issues are not handled in the open, as it were, but in the closed spaces of selection committees and elsewhere. The bottom line is always the scientific credibility

of the *methodology*, an issue that in these places is seldom open for debate. The result is that innovations are assimilated to the existing methodologies. Many of the pioneers of the new cognitivism have, themselves, come to regret this loss of nerve, 'cognitive psychology lost out to the received view, with its operational and reductionistic methods ... the old won out over the new' (Garner, 1999, p. 21; see also Jenkins, 1986).

The fate of the ambitions of the pioneers of the new cognitivism seems to confirm Danziger's claim that the 'objects' of psychological discourse are shaped by the existing research practices 'so as to fit the Procrustean bed of a very limited range of allowable procedures' (Danziger, 1996, p. 17). However, the objectivist ideal of methodology that modern cognitivism took over from behaviourism was never a reality, but a creation of Watson's lively imagination. The result is that psychologists remain committed to an unconscious *theory* of psychological methodology that implies mind-body dualism.

15.4 Theory of mind

I will finish with an example of where the *unconscious* theory implicit in modern cognitive psychology's three-stage seems to have finally come into the light of day as *explicit* theory – but where the fundamental premises of the theory have, yet again, remained as *unconscious* as ever.

'Theory of Mind' was a bad mistake just waiting to happen. If psychologists can only make sense of other people in an indirect way, surely doesn't this have to be the case for how 'other people' (non-psychologists) make sense of other people? By the 1980s, psychologists eventually projected their *unconscious* theory of psychological methodology onto the people they were studying as an apparently *explicit* theory of how people, in general, make sense of other people.

So, according to Theory of Mind (ToMism), *everyone* is now supposed to be a psychologist. In our encounters with other people, we all necessarily begin as reductionistic, methodological behaviourists, since all we can observe is their 'mere' behaviour. In order for other people to make sense to us, we must engage in a separate stage of theorizing (or quasi-theorizing) in order to bridge the gulf between behaviour and mind. The original version of ToMism, 'Theory theory' claimed that we *literally* engage in theorizing on the basis of the limited available data. This version has been followed by others, for example, innate neural modules which conveniently do the theorizing for us, or simulation theory which, in its original form, claimed that when observing

the behaviour of another person, we try to imagine how we would feel under similar circumstances and then project those feelings onto that other person (for a critical overview of ToMism see Leudar and Costall, 2009).

Since the 1980s, an astonishing amount of research and theorizing has gone into determining when and how young children solve this 'problem of other minds'. However, the way this problem has been formulated has simply been taken for granted, namely, how we can possibly understand other people psychologically, given (1) that the only available evidence available to us are mere 'movements in space' (Meltzoff, Gopnik, and Repacholi, 1999, p. 17), and (2) that mental states are 'peculiarly private' (Leslie, 1999, p. 576). In fact, these two crucial theoretical assumptions are not regarded as *theory* at all, even though they had been consistently questioned within the philosophy of mind long before the rise of Theory of Mind or even Modern Cognitivism.

The ToMists also take for granted that 'the problem of other minds' – *as they have formulated it* – must have a solution since people do, after all, make sense of one another. Yet, *their* problem – given the above taken-for-granted assumptions – is none other than the traditional 'problem of other minds' and *that* problem is, in principle, insoluble. It is hardly surprising, therefore, that the theoretical 'solutions' that have been proposed never spell out *exactly* how we are all supposed to bridge the gap between 'mere' behaviour and mind. For example, natural selection (a favourite *deus ex machina* of modern cognitivist theory) is somehow supposed to construct innate theory of mind neurological modules, even though no account is given of how natural selection itself could possibly detect and hence differentiate 'peculiarly private' mental states. And the Theory Theorists (e.g. Gopnik, 2003, p. 245) see no need themselves to explain how young children are actually supposed to bridge the gulf between 'mere' behaviour and mind. They do not regard this as any of *their* theoretical business. As Ben Bradley was quick to take note:

> TOM is the culmination of the movement in modern psychology to endow the young child with ever greater intellectual powers ... The child is no longer just competent, perceptually, socially and cognitively. It is now a theoretician. In two years it has solved problems that have stumped philosophers for two millennia. (Bradley, 1993, p. 507)

The ToMists, along with modern cognitivists in general, are right to insist that they bear no intellectual allegiance to Descartes and his *ontological* dualism of mind and body. This is why they get *so* angry when

they are accused of not just being conceptually confused but also of being *dualists*. The power of unconscious theory is that it *is* unconscious. In this case, cognitive psychologists managed to reinvent the traditional problem of other minds through their unwitting acceptance of the *methodological* dualism of mind and behaviour contained in Watson's historical myths. Consequently, the most fundamental theoretical assumption of modern psychology is now widely mistaken for established historical fact, and hence in no need of any critical analysis.

15.5 Conclusion

In this chapter, I have invoked psychoanalytic terms such as 'unconscious', 'denial' and 'condensation'. I am not proposing, however, that cognitive psychologists should be subjected to psychoanalysis rather than conceptual analysis – well, not *all* of them. I am not setting myself up in opposition to conceptual analysis. For example, Bennett and Hacker's *Philosophical Foundations of Neuroscience* (2003) is an impressive corrective to the confusions surrounding the latest phase of Cognitivism, Cognitive Neuroscience, where the localization of psychological functions in the brain has become a displacement activity for avoiding the serious scientific issues. However, as I have argued in this chapter, many of the most fundamental conceptual issues in modern cognitivism are *implicit* in the language and practice of psychology. Unconscious theory has a lot to answer for. Psychology could be described as the study of people who probably do not need to be studied, by people who most certainly do. Getting to grips with the deep confusions that beset modern Psychology will surely require a much closer examination of what psychologists are actually doing when they are *doing* psychology.

References

B. J. Baars (ed.) (1986) *The Cognitive Revolution in Psychology* (New York: Guilford Press).

F. C. Bartlett (1932) *Remembering* (Cambridge: Cambridge University Press).

M. R. Bennett and P. M. S. Hacker (2003) *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).

B. Bradley (1993) 'A Serpent's Guide to Children's "Theories of Mind"', *Theory and Psychology*, 3, 497–521.

M. Brewer and M. Hewstone (2004) 'Introduction' in M. Brewer and M. Hewstone (eds) *Social Cognition* (Oxford: Blackwell), xi–xii.

A. Brock (1991) 'Imageless Thought or Stimulus Error? The Social Construction of Private Experience' in W. R. Woodward and R. S. Cohen (eds) *World Views and Scientific Discipline Formation* (Dordrecht: Kluwer Academic Publishers), 97–106.

J. S. Bruner, J. J. Goodnow, and G. A. Austin (1956) *A Study of Thinking* (New York: John Wiley).
N. R. Carlson, C. D. Heth, H. Miller, J. W. Donahue, W. Buskist, and G. N. Martin (2006) *Psychology: The Science of Behaviour*, 6th edn (Boston: Pearson).
A. Costall (1991) 'Graceful Degradation: Cognitivism and the Metaphors of the Computer' in A. Still and A. Costall (eds) *Against Cognitivism: Alternative Foundations for Cognitive Psychology* (Hemel Hempstead: Harvester-Wheatsheaf), 151–70.
—— (2006) '"Introspectionism" and the Mythical Origins of Scientific Psychology', *Consciousness and Cognition*, 15, 634–54.
—— (2007) 'Bringing the Body Back to Life: James Gibson's Ecology of Agency' in J. Zlatev, T. Ziemke, R. Frank, and R. Dirven (eds) *Body, Language and Mind: Vol. 1: Embodiment* (The Hague: de Gruyter), 241–70.
K. Danziger (1990) *Constructing the Subject: Historical Origins of Psychological Research* (Cambridge: Cambridge University Press).
K. Danziger (1996) 'The Practice of Psychological Discourse' in C. F. Graumann and K. J. Gergen (eds) *Historical Dimensions of Psychological Discourse* (Cambridge: Cambridge University Press), 17–35.
J. Dewey (1896) 'The Reflex Arc Concept in Psychology', *Psychological Review*, 3, 357–70.
—— (1914/1977) 'Psychological Doctrine and Philosophical Teaching' in S. Morgenbesser (ed.) *Dewey and His Critics* (New York: Journal of Philosophy, Inc.), 439–45. [First published in the *Journal of Philosophy Psychology and Scientific Methods*, 1914, 11(19)].
H. Dreyfus (1995) 'Cognitivism Abandoned' in P. Baumgartner and S. Payr (eds) *Speaking Minds: Interview with Twenty Eminent Cognitive Scientists* (Princeton, NJ: Princeton University Press), 70–83.
A. K. Ericsson and H. A. Simon (1984) *Protocol Analysis: Verbal Reports as Data*. (Cambridge, MA: MIT Press).
W. R. Garner (1999) 'Reductionism Reduced: Review of "Toward a New Behaviourism: The Case against Perceptual Reductionism" by William R. Uttal', *Contemporary Psychology*, 44, 20–21.
G. Gigerenzer and D. G.. Goldstein (1996) 'Mind as Computer: Birth of a Metaphor', *Creativity Research Journal*, 9, 131–44.
A. Gopnik (2003) 'The Theory as an Alternative to the Innateness Hypothesis' in L. M. Antony and N. Hornstein (eds) *Chomsky and His Critics* (Oxford: Blackwell), 238–54.
D. O. Hebb (1966) *Textbook of Psychology*, 2nd edn (New York: Saunders).
E. Heidbreder (1933) *Seven Psychologies* (New York: Century).
C. L. Hull (1943) *Principles of Behaviour* (New York: D. Appleton-Century Co.).
E. Jacobson (1973) 'Electrophysiology of Mental Activities and Introduction to the Psychological Process of Thinking' in F. J. McGuigan and R. A. Schoonover (eds) *Psychophysiology of Thinking* (New York: Academic Press).
J. J. Jenkins (1986) 'Interview with James J. Jenkins' in B. J. Baars (ed.) *The Cognitive Revolution in Psychology* (New York: Guilford Press), 239–52.
P. Johnson-Laird (1988) *The Computer and the Mind* (Cambridge, MA: Cambridge University Press).
R. F. Kitchener (1977) 'Behaviour and Behaviourism', *Behaviourism*, 5, 11–71.

S. Koch (1981) 'The Nature and Limits of Psychological Knowledge: Lessons of a Century qua "Science"', *American Psychologist*, 36, 257–69.
K. Kroker (2003) 'The Progress of Introspection in America, 1896–1938', *Studies in the History and Philosophy of Biology and Biomedical Science*, 34, 77–108.
A. Leslie (1999) 'Mind, Child's Theory of' in *Concise Routledge Encyclopaedia of Philosophy* (London: Routledge), 575–6.
I. Leudar and A. Costall (eds) (2009) *Against Theory of Mind* (London: Palgrave Macmillan).
A. N. Meltzoff, A. Gopnik, and B. M. Repacholi (1999) 'Toddlers' Understanding of Intentions, Desires and Emotions: Explorations of the Dark Ages' in P. D. Zelazo, J. W. Astington, and D. R. Olson (eds) *Developing Theories of Intention* (Mahwah, NJ: Erlbaum), 17–41.
C. K. Monson and R. T. Hurlburt (1993) 'A Comment to Suspend the Introspection Controversy: Introspecting Subjects did Agree about Imageless Thought' in R. T. Hurlburt (ed.) *Sampling Inner Experience in Disturbed Affect* (New York: Plenum), 15–26.
A. Newell, J. C. Shaw, and H. A. Simon (1958/1991) 'Elements of a Theory of Human Problem Solving' in A. B. Schoedinger (ed.) *Introduction to Metaphysics: The Fundamental Questions* (Buffalo, NY: Prometheus Books), 385–406. [Article originally published in 1958].
M. J. Packer (2008) 'Is Vygotsky Relevant? Vygotsky's Marxist Psychology', *Mind, Culture, and Activity*, 15, 8–31.
Z. Pylyshyn (1986) *Computation and Cognition* (Cambridge, MA: MIT Press).
R. E. Shaw (2003) 'The Agent-Environment Interface: Simon's Indirect or Gibson's Direct Coupling', *Ecological Psychology*, 15, 37–106.
H. A. Simon (1991) *Models of My Life* (New York: Basic Books).
E. L. Thorndike (1915) 'Watson's "Behaviour"', *Journal of Animal Behaviour*, *5,* 462–7.
E. C. Tolman (1932) *Purposive Behaviour in Animals and Men* (New York: Appleton-Century-Crofts).
S. Toulmin (1969) 'Ludwig Wittgenstein', *Encounter*, 32, 58–71.
P. J. van Strien and E. Faas (2004) 'How Otto Selz Became a Forerunner of the Cognitive Revolution' in T. C. Dalton and R. B. Evans (eds) *The Life Cycle of Psychological Ideas: Understanding Prominence and the Dynamics of Intellectual Change* (New York: Kluwer Academic/Plenum Publishers), 175–201.
M. F. Washburn (1915) 'Review of *Behaviour: An Introduction to Comparative Psychology* by John B. Watson', *Philosophical Review,* 24, 210–13.
J. B. Watson (1913) 'Psychology as the Behaviourist Views It', *Psychological Review*, 20, 158–77.
—— (1914) *Behaviour: An Introduction to Comparative Psychology* (New York: Holt).
—— (1920) 'Is thinking merely the action of language mechanisms?' *British Journal of Psychology*, 11, 87–104.
—— (1924) *Psychology from the Standpoint of a Behaviourist*, 2nd edn (Philadelphia: J. B. Lippincott Company).
—— (1925) *Behaviourism* (London: Kegan Paul, Trech, Trubner, and Co).
R. S. Woodworth (1918) *Dynamic Psychology* (New York: Columbia University Press).
R. S. Woodworth (1931) *Contemporary Schools of Psychology* (London: Methuen and Co.).

Index