Current Controversies in Philosophy

CURRENT CONTROVERSIES IN PHILOSOPHY OF COGNITIVE SCIENCE

Edited by
Adam J. Lerner, Simon Cullen,
and Sarah-Jane Leslie

ROUTLEDGE

Current Controversies in Philosophy of Cognitive Science

Cognitive science is the study of minds and mental processes. Psychology, neuroscience, computer science, and philosophy, among other subdisciplines, contribute to this study. In this volume, leading researchers debate five core questions in the philosophy of cognitive science:

- Is an innate Universal Grammar required to explain our linguistic capacities?
- Are concepts innate or learned?
- What role do our bodies play in cognition?
- Can neuroscience help us understand the mind?
- Can cognitive science help us understand human morality?

For each topic, the volume provides two essays, each advocating for an opposing approach. The editors provide study questions and suggested readings for each topic, helping to make the volume accessible to readers who are new to the debates.

Adam J. Lerner is Assistant Professor/Faculty Fellow at the New York University Center for Bioethics. He completed his PhD in Philosophy at Princeton University in 2018 and he works on issues in ethics, metaethics, moral psychology, and the philosophy of mind.

Simon Cullen is Assistant Teaching Professor of Philosophy at Carnegie Mellon University. He earned his PhD in Philosophy at Princeton University in 2015 and was a postdoctoral research fellow at Princeton Neuroscience Institute in 2017. His work focuses on the folk concept of self, especially the notion of a "true self" and its theoretical and normative implications; developing empirical methods to advance experimental philosophy and other areas of social scientific inquiry; and helping people improve at open-minded analytical reasoning and communication.

Sarah-Jane Leslie is the Class of 1943 Professor of Philosophy and Dean of the Graduate School at Princeton University. She is the author of numerous articles in philosophy and psychology, published in journals such as *Science*, *Proceedings of the National Academy of Sciences*, *Philosophical Review*, and *Noûs*.

Current Controversies in Philosophy

Series Editor: John Turri, *University of Waterloo*

In venerable Socratic fashion, philosophy proceeds best through reasoned conversation. Current Controversies in Philosophy provides short, accessible volumes that cast a spotlight on ongoing central philosophical conversations. In each book, pairs of experts debate four or five key issues of contemporary concern, setting the stage for students, teachers and researchers to join the discussion. Short chapter descriptions precede each chapter, and an annotated bibliography and suggestions for further reading conclude each controversy. In addition, each volume includes both a general introduction and a supplemental guide to further controversies. Combining timely debates with useful pedagogical aids allows the volumes to serve as clear and detailed snapshots, for all levels of readers, of some the most exciting work happening in philosophy today.

Published Volumes in the Series:

Current Controversies in Bioethics
Edited by S. Matthew Liao and Collin O'Neil

Current Controversies in Philosophy of Film
Edited by Katherine Thomson-Jones

Current Controversies in Metaphysics
Edited by Elizabeth Barnes

Current Controversies in Values and Science
Edited by Kevin C. Elliott and Daniel Steel

Current Controversies in Philosophy of Religion
Edited by Paul Draper

Current Controversies in Philosophy of Cognitive Science
Edited by Adam J. Lerner, Simon Cullen, and Sarah-Jane Leslie

For more information about this series, please visit:
www.routledge.com/Current-Controversies-in-Philosophy/book-series/CCIP

Current Controversies in Philosophy of Cognitive Science

Edited by
Adam J. Lerner, Simon Cullen, and Sarah-Jane Leslie

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2020
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2020 Taylor & Francis

The right of Adam J. Lerner, Simon Cullen and Sarah-Jane Leslie to be identified as authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-1-138-85800-8 (hbk)
ISBN: 978-1-003-02627-3 (ebk)

Typeset in Bembo
by Newgen Publishing UK

Contents

Notes on Contributors vii

Introduction 1

PART I
Is There a Universal Grammar? 11

1 Universal Grammar 13
PAUL PIETROSKI AND NORBERT HORNSTEIN

2 Waiting for Universal Grammar 29
GEOFFREY K. PULLUM

Further Readings for Part I 44

Study Questions for Part I 46

PART II
Are All Concepts Learned? 47

3 Beyond Origins: Developmental Pathways and the Dynamics of Brain Networks 49
LINDA B. SMITH, LISA BYRGE, AND OLAF SPORNS

4 The Metaphysics of Developing Cognitive Systems: Why the Brain Cannot Replace the Mind 63
MARK FEDYK AND FEI XU

Further Readings for Part II 83

Study Questions for Part II 84

PART III
What Is the Role of the Body in Cognition? 85

5 Embodied Cognition and the Neural Reuse Hypothesis 87
JULIAN KIVERSTEIN

6 Rehashing Embodied Cognition and the Neural Reuse Hypothesis 108
FRED ADAMS

Further Readings for Part III 117

Study Questions for Part III 118

PART IV
How Should Neuroscience Inform the Study of Cognition? 119

7 Is Cognitive Neuroscience an Oxymoron? 121
FIERY CUSHMAN

8 On the Primacy of Behavioral Research for Understanding the Brain 134
YAEL NIV

Further Readings for Part IV 150

Study Questions for Part IV 152

PART V
What Can Cognitive Science Teach Us About Ethics? 153

9 The Ethical Significance of Cognitive Science 155
VICTOR KUMAR

10 Putting the "Social" Back in Social Psychology 174
COLIN KLEIN

Further Readings for Part V 188

Study Questions for Part V 190

Index 191

Contributors

Fred Adams is Professor of Linguistics and Cognitive Science at the University of Delaware. He is also Professor of Philosophy there. He has published over 150 articles on philosophy and cognitive science, in addition to his jointly published book *Bounds of Cognition* (2007) and Introduction to the *Philosophy of Psychology* (2018).

Lisa Byrge is Postdoctoral Researcher at Indiana University. Her research focuses on detecting and understanding individual differences in brain function using fMRI.

Fiery Cushman is John L. Loeb Associate Professor of Social Sciences in the Department of Psychology at Harvard University. His research focuses on value-guided decision making and, especially, moral judgment and decision making.

Mark Fedyk is Associate Professor in the Bioethics Program of the UC Davis School of Medicine and the Betty Irene Moore School of Nursing. He is the author of *The Social Turn in Moral Psychology* (2017) as well as a variety of articles and chapters that examine how normative phenomena can be studied using scientific methods.

Norbert Hornstein teaches linguistics at the University of Maryland. He has written several books on syntactic theory including *A Theory of Syntax* (2009) and *Move!* (2001).

Julian Kiverstein is Senior Researcher at the University of Amsterdam Medical Centre based in the Psychiatry Department. He has published extensively in the philosophy of embodied cognition, including the monograph *Extended Consciousness and Predictive Processing: A Third-Wave View*, co-authored with Michael Kirchhoff (2019).

Colin Klein is Associate Professor of Philosophy at the Australian National University. His work focuses on philosophy of neuroscience, pain perception (*What the Body Commands*, 2015), and online trust. He is also a CI on the ANU Humanizing Machine Intelligence project, where he explores the philosophical and ethical implications of machine learning algorithms.

Victor Kumar is Assistant Professor at Boston University. He works mainly at the intersection of ethics and cognitive science. His published work can be found in *Ethics*, *Noûs*, and *Philosophers' Imprint*. In recent years he has written about moral learning, moral luck, and moral disgust. He is currently finishing a book with Richmond Campbell about moral evolution and moral progress.

Yael Niv is Professor in the Psychology Department at Princeton University, and at the Princeton Neuroscience Institute. Her lab's research focuses on the neural and computational processes underlying reinforcement learning and decision making. She is co-founder and co-director of the Rutgers–Princeton Center for Computational Cognitive Neuropsychiatry.

Paul Pietroski is Distinguished Professor of Philosophy at Rutgers University. His research addresses questions concerning linguistic meaning and its relation to cognition. His most recent book is *Conjoining Meanings: Semantics without Truth Values* (2018). He has held visiting positions at Harvard University and the École Normale, Paris.

Geoffrey K. Pullum is Professor of General Linguistics in the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh. He is a Fellow of the British Academy and of the Linguistic Society of America. He has published widely in theoretical linguistics, English grammar, and philosophy of linguistics.

Linda B. Smith is Distinguished Professor of Psychological and Brain Sciences and Chancellor's Professor of Psychological and Brain Sciences at Indiana University, Bloomington.

Olaf Sporns is Distinguished Professor in the Department of Psychological and Brain Sciences at Indiana University in Bloomington. His main research area is theoretical and computational neuroscience, with a focus on complex brain networks. In addition to numerous peer-reviewed journal articles, he has authored two books, *Networks of the Brain* and *Discovering the Human Connectome*.

Fei Xu is Professor of Psychology at the University of California, Berkeley. Her research focuses on cognitive and language development from infancy to middle childhood. She has published more than 100 journal articles and book chapters. In recent years, she and her collaborators have developed a *rational constructivist* theory of cognitive development, with a strong emphasis on characterizing mechanisms of learning such as language/symbol learning, Bayesian inductive learning, and constructive thinking.

Introduction

Cognitive science is the study of minds and mental processes. Psychology, neuroscience, computer science, and philosophy, among other subdisciplines, all contribute to this study. Consequently, work done under the heading "philosophy of cognitive science" varies significantly. Some of this work could be more accurately described as philosophy *in* cognitive science, wherein philosophical methods are used to advance first-order debates in cognitive science that have a distinctly philosophical flavor (e.g., concerning cognitive architecture, the innateness of concepts, the ability to think about other minds). Some of what gets called "philosophy of cognitive science" could be more accurately described as philosophy *with* cognitive science or even cognitive science of philosophy, in which researchers draw on results from cognitive science to advance a wide variety of questions in philosophy concerning not just the metaphysics of mind, but also epistemology, ethics, philosophy of language, and philosophical methodology in general. The rest of what gets called "philosophy of cognitive science" is most accurately described as philosophy of cognitive science, which addresses the methodological and general ontological presuppositions of cognitive science (e.g., concerning the nature of representation, the relationship between modeling and ontology, what explanation requires, the possibility of reduction, standards of theory choice).

But these distinct strands of work in philosophy of cognitive science are not entirely independent. Indeed, most chapters in this volume address both first-order questions about cognition and second-order (methodological) questions about how best to study cognition. The authors argue that what we say about the first-order questions should inform our methodological views, and vice versa. As will become evident, a dominant thread running throughout the volume involves the rippling implications of new empirical methods for addressing questions that have been traditionally considered *a priori* or best pursued using older empirical methods.

Part I

The volume begins with two cutting-edge contributions to perhaps the 20th century's most contentious debate in philosophy of cognitive science: the debate over Universal Grammar. In their contribution, Paul Pietroski and Norbert

Hornstein offer a defense of the view that humans possess a Universal Grammar, an innate capacity to acquire procedures for assigning meanings to pronunciations. On this Chomskyan view, the capacity in question either provides a "common template" that experience "fills in" to arrive at a procedure, or else it comes with a limited set of procedures as options, one of which is selected on the basis of exposure to local speech. Pietroski and Hornstein posit Universal Grammar because, they argue, it explains various features of our linguistic capacities.

First, it explains linguistic creativity: Humans can generate and comprehend indefinitely many strings of words they have never encountered before. Second, it explains universal availability: Given the right experience, any human child can learn any language that any other human child can learn. Moreover, Pietroski and Hornstein argue that we must invoke Universal Grammar to explain why there are certain *constraints* on the ways in which humans can be linguistically creative. For example, certain meanings (e.g., *The goat is eager to be consumed*) cannot be assigned to certain word-strings (e.g., "the goat is eager to eat"), despite superficial resemblances between those word-strings and other word-strings (e.g., "the goat is ready to eat") where similar meanings *can* be assigned (e.g., *The goat is fit to be consumed*). Such cases of "limited homophony" suggest that humans acquire *restrictive* procedures for assigning meanings to word-strings rather than more *permissive* procedures that would allow for word-strings to be more ambiguous than they are.

Pietroski and Hornstein argue that the speech-based evidence available to each child (i.e., the particular, idiosyncratic set of utterances that each child hears) does not support restrictive procedures over more permissive ones. Moreover, such evidence does not support the specific restrictive procedures that children actually arrive at over other possible restrictive procedures. So, to explain why humans arrive at only a small set of restrictive procedures based on such limited evidence, we must assume humans have a Universal Grammar which in some way takes many possible procedures off the table.

In a critical response to Pietroski and Hornstein, Geoffrey Pullum questions both the assumption that human languages are similar enough to be explained by a Universal Grammar, as well as the assumption that children receive an impoverished diet of speech that could not allow them to acquire such restrictive procedures. On the first assumption, he points out that learning the meanings of individual words requires attention to many properties that vary between languages and that are ignored by Universal Grammar. Such properties include phonology, inflection, derivational relationships, some syntactic properties, literal meaning, conventional implicatures, and overtones and associations. On the second assumption, he accuses Pietroski and Hornstein of neglecting contemporary work on inductive learning that can explain how children acquire languages without the help of a Universal Grammar.

First, Pullum argues that Pietroski and Hornstein neglect research showing that young children hear over a million utterances per year. Second, he argues that they ignore research showing how languages could have evolved to be learnable for beings with our general capacities. Pullum argues that this provides a better explanation of the similarities between languages than does Universal Grammar. Third,

he argues that they ignore progress in Bayesian statistical learning that provides an explanation of how it could be rational to assume certain expressions are ungrammatical despite never hearing them. Finally, he argues that they ignore the contribution that *context* and *meaning* play in language acquisition. For example, he agrees with Pietroski and Hornstein that "Karen saw her husband with another woman" cannot be interpreted in such a way that it could be true in virtue of Karen seeing her husband while finding herself in the company of her friend Maureen. But while Pietroski and Hornstein argue that the unavailability of this interpretation reflects constraints imposed by Universal Grammar—presumably that *with*-phrases at the end of a transitive clause cannot modify the subject—Pullum argues that the unavailability of this interpretation instead reflects our ability to take into account context when assigning plausible interpretations. In support of this claim, he points out that *with*-phrases can modify the subject in other contexts: "Karen watched the eclipse with another woman." The reason that the *with*-clause clearly modifies the subject in this case is that taking it to modify the direct object would require us to implausibly assume that the speaker is attempting to convey a bizarre meaning, a meaning that could be true in virtue of Karen witnessing a scene in which Maureen is in the company of an eclipse.

Part II

The debates surrounding a putative Universal Grammar speak primarily to the syntax of language and the structure of human cognition. An even more heated and protracted debate concerns the origin of cognitive content: Are concepts innate or learned? Linda Smith, Lisa Byrge, and Olaf Sporns argue that this traditional question is not well-posed, both because the notions *innateness* and *learning* are irredeemably vague, and also because contemporary cognitive science has no need to posit the existence of *concepts* to explain the development of human cognition. In lieu of discussing the origins of concepts, Smith and colleagues review a large body of work revealing that changes in the child's body, behavior, and environment over the course of development have a crucial role to play in the development of the human brain. They argue that by situating the brain within this *brain–behavior–environment network*, we can better understand how short-term brain dynamics arise and influence long-term changes in functional and structural connectivity. Because the development of human cognition involves multiple factors influencing one another in complex ways over multiple time scales, Smith and colleagues conclude that contemporary cognitive science renders moot traditional questions about the innateness of human concepts.

Following Smith and colleagues, Mark Fedyk and Fei Xu endorse the view that the brain, body, and environment influence one another in complex ways over the course of cognitive development. But unlike Smith and colleagues, they take this "dynamic systems" view to be compatible with the existence of *concepts* that are *innate* in an important sense. For example, they follow Susan Carey in thinking that young children have innate concepts such as *object*, *agent*, *number*, and perhaps *cause*. To say these concepts are innate is to say, roughly, that they are *essential to the development* of at least one simple cognitive system. In the case of concepts such

as *object*, *agent,* and *number*, the simple cognitive systems in question are dedicated input analyzers, which take as input non-conceptual information and produce richer, conceptual representations as output. Fedyk and Xu argue further that a version of Smith's dynamic systems view that is supplemented with such innate concepts can explain more data than relying solely on Smith's more restricted view. In support of this claim, they cite evidence both that children and adults regularly make inferences that follow basic principles of logic and probability, among other rational patterns of inference, and that such inferences have the characteristics that David Marr argued a process must have to qualify as *psychological computation*. They argue that psychological computation of this sort cannot be adequately explained without positing a *sui generis* cognitive system with innate concepts over and above the brain-behavior-environment network. Drawing on recent work in philosophy of science on the metaphysics of mechanisms, Fedyk and Xu provide a detailed account of how a *sui generis* cognitive system could be co-instantiated with dynamic systems like the brain-behavior-environment network.

Part III

From debates about the origins of concepts, we turn to a more recent dispute about the nature of representation and cognition generally: In what respects and to what extent is human cognition *embodied*, and what implications does this have for cognitive science? In his contribution, Julian Kiverstein describes two recent approaches to embodied cognition that answer these questions, both of which draw on work on "neural reuse" from Michael Anderson. The first "moderate" account comes from Alvin Goldman, while the second comes from Anderson himself. On Goldman's account, human cognition is embodied in virtue of and to the extent that it makes use of "bodily formatted representations" ("B-formatted representations")—representations that originally evolved to represent states of the body—to execute novel cognitive tasks. According to Goldman, human cognition is more thoroughly embodied in this sense than cognitive scientists have traditionally recognized, but recognizing this poses no threat to traditional views in cognitive science. On traditional views, human cognition consists in computational processes operating over inner mental representations, and embodied cognition simply highlights the extent to which B-formatted representations serve as input to these processes. On Anderson's account, countenancing embodied cognition is far more radical. On this view, embodied cognition consists in a complex web of interacting components, including not just activity in sensorimotor regions of the brain, but the body itself and its interaction with the surrounding environment. Moreover, Anderson argues that higher cognition is typically—if not always—embodied cognition of this more radical sort. Unlike Goldman's view, Anderson's view appears to demand a complete overhaul of cognitive science, including the rejection of the traditional view that cognition primarily consists in computational processes defined over inner mental representations. In its place, cognition should be studied not as a phenomenon primarily located in the brain, but as a process constituted across the brain, the body, and the surrounding environment.

Both Goldman and Anderson support their respective views by drawing on the empirical literature on "neural reuse." This literature includes Anderson's own meta-analytic work revealing many discrete brain areas are activated across a wide variety of tasks, as well as work revealing both facilitative and inhibitory effects on people's cognitive performance while they perform motor behaviors that rely on the same neural circuitry in a similar or different way (e.g., recalling a greater number of positive memories when moving marbles in a positive, upwards direction; judging the meaning of a sentence more slowly when reporting their judgment required moving in a way opposed to the action described in the sentence). But Kiverstein argues that empirical work on neural reuse supports Anderson's more radical view of embodied cognition against Goldman's more moderate view. To bolster this interpretation of the evidence, Kiverstein first defends a "mutualist" view of evolution on which, instead of organisms evolving in response to selection pressures posed by a more or less stable environment, organisms and their environments have a far more intimate and reciprocal influence on one another. He then argues that, when viewed from this mutualist perspective, Anderson's radical view of embodied cognition—a view that assigns the body itself and sensorimotor activation a constitutive role in human cognition—provides a better explanation of the data than Goldman's more moderate view.

In his reply, Fred Adams claims to find nothing in the data or Kiverstein's arguments that supports Anderson's more radical view over Goldman's more conservative view. First, he distinguishes between "weak" and "strong" forms of embodiment. On "weak" forms of embodiment, neither the body nor "reused" sensorimotor neural activity plays a constitutive role in cognition. At most, sensorimotor neural activity provides representations for central cognitive processes to take as input for further computational processing. On "strong" forms of embodiment, sensorimotor neural activity—and perhaps even the body itself—can play a constitutive role in cognition. Unlike weak forms of embodiment, strong forms of embodiment are not compatible with the traditional view of cognition, on which cognition is—as Susan Hurley once put it—"sandwiched" between perception and action. Having distinguished between these forms of embodiment, Adams argues that Goldman merely embraces a weak form of embodiment, and that Anderson himself does not clearly embrace a strong view.

Against Kiverstein, Adams argues that Goldman is right to embrace only a weak form of embodiment. He argues that none of the data in question shows that sensorimotor neural activity or the body are constitutively implicated in cognitive processes in the way Kiverstein claims. In each case, the explanation of the findings that rely on a strong view of embodiment are no stronger than the findings that rely on a weak view. His key claim is that the body and sensorimotor activity may merely causally influence rather than partially constitute cognitive processing. He likewise distinguishes between "causal mutualism" and "constitutive mutualism" and claims that the evolutionary considerations that Kiverstein adduces are compatible with causal mutualism, which is in turn compatible with the "sandwich" model of cognition. He concludes that while Goldman's view of embodied cognition may make embodied cognition less interesting than Kiverstein believes it to be, Goldman recognizes this and is right to remain unconcerned.

Part IV

Many current controversies in philosophy of cognitive science concern the extent to which empirical work in cognitive science can address philosophical questions, and the extent to which different branches of cognitive science can inform each other. In their contributions, Fiery Cushman and Yael Niv take up a question of this last type: To what extent can neuroscience inform our understanding of the mind? While Cushman argues that neuroscientific theories may do little to advance our understanding of the mind, neuroscientific *methods* can play an important role. He supports this optimistic conclusion in three distinct ways. First, he provides an *a priori* analysis of how—given that the brain is the physical substrate of the mind—both neuroscientific manipulation (e.g., transcranial magnetic or electrical stimulation, lesion, optogenetics) and measurement (e.g., fMRI, EEG, single-unit recording) of the brain could in principle constrain models of human cognition. He then demonstrates how this in-principle possibility can be implemented in practice by describing a case study in which single-unit recordings of retinal ganglion cells led to significant breakthroughs in computational models of visual processing. Finally, to show that such cases of neuroscience-induced progress of cognitive science may not be few and far between, he shares data showing that articles published in the journal *Cognition* that cited at least one neuroscience article were cited more often than articles that did not—a statistically significant correlation that holds even when controlling for several potential confounds (e.g., year of publication, number of citations in each article). He argues that a plausible interpretation of this finding is that such articles are in fact more influential, and more influential because they better achieve the goals of cognitive science. While Cushman recognizes that alternative hypotheses remain that are consistent with the data, he argues that combining this data with the *a priori* analysis and case study of visual processing results in a weighty body of evidence in favor of the view that neuroscience has an important role to play in cognitive science. Even if neuroscience may not be *necessary* or *sufficient* for the advancement of cognitive science, it can play an important role in expediting our understanding of human cognition.

Yael Niv agrees with Cushman that neuroscientific methods *can* advance research into human cognition. And while Cushman recognizes that there may be limits to what neuroscience can contribute to cognitive research, Niv places much more emphasis on these limits. Indeed, despite her own use of neuroscientific methods, Niv argues that critical voices accusing neuroscience of being overvalued and receiving undue support—from funding bodies, academic researchers, and journals—are well-founded.

First, she argues that, except in rare cases (e.g., the visual processing studies described by Cushman), well-designed experiments utilizing behavioral methods are necessary for the advancement of cognitive research. She supports these claims by drawing on many examples of progress in cognitive science, including breakthroughs in understanding prediction error and reinforcement learning, the structure and format of working memory, and the control of attention. She argues further that, except in rare cases (e.g., lesion studies on memory, future

work on motivational systems), behavioral methods are far more important for understanding human cognition than neuroscientific methods. Moreover, she provides evidence that understanding the neural implementation of cognitive processes—a task well-suited for study by neuroscientific methods, if any is—depends on behavioral research, and that a great deal of research into questions of implementation (e.g., on the structure of working memory) can be successfully conducted entirely with cleverly designed behavioral research. While such research often depends on correlational measures, she contends that adequate design can help us avoid conflating correlation with causation and help us make progress much more quickly and affordably than experimental interventions on the brain. Although Niv believes both behavioral research and neuroscientific research have important roles to play in cognitive science, she concludes with a call to recognize the importance and primacy of behavioral research.

Part V

While Cushman and Niv engage with the widespread assumption that the methods of neuroscience provide the best way to achieve the goals of cognitive science, the final pair of contributions to the volume engage with the widespread assumption that the methods of cognitive science are all but irrelevant to moral philosophy. Against the philosophically dominant view that cognitive science has nothing to contribute to moral philosophy, Kumar argues that cognitive science provides crucial input to at least three distinct projects within moral philosophy.

First, he argues that cognitive science is especially well-placed to help us make progress on two central metaethical questions: Are moral judgments cognitive or non-cognitive states? And what is the link between moral judgment and motivation? Central to this argument is the assumption, which Kumar defends at length elsewhere, that moral judgments are *natural psychological kinds* whose nature is most aptly investigated using the methods of cognitive science. Because the methods of cognitive science can shed light on the causal/explanatory role of moral judgment in human psychology, they can shed light on the question of whether moral judgments are cognitive or non-cognitive states. Kumar then argues that taking this approach favors his own "hybrid" theory of moral judgment over other empirically motivated theories, such as Jesse Prinz's "emotionism." Turning to the question of whether agents' moral judgments necessarily or contingently motivate them to act accordingly, Kumar argues that, by pursuing his empirically oriented approach to metaethics, we can make progress. He argues that the cognitive science of moral judgment supports the view that moral motivation can be both constitutive of moral judgment and yet only contingently associated with it.

Next, Kumar argues that cognitive science can supply the ingredients for arguments that can debunk intuitions that serve as input to moral theorizing, thereby having an important downstream influence on first-order moral theorizing. Setting evolutionary debunking arguments aside as less successful than proponents believe, Kumar considers several ways that empirical investigation into more proximal causes of our moral judgments could debunk them, but he argues that these can be successful only insofar as their targets remain narrow. He

contends that more ambitious arguments—including Walter Sinnott-Armstrong's argument for moral skepticism from framing effects, Joshua Greene's arguments against deontology from heuristics and biases, and Daniel Kelly's arguments against disgust-based beliefs—inevitably rely on empirical speculation that goes beyond what is justified by the available data. Drawing on work with Richmond Campbell, Kumar then offers a recipe for how to develop selective debunking arguments that remain compatible with empirical findings.

Finally, drawing on non-ideal political theory as a model, Kumar advocates for an empirically informed non-ideal moral theory. Such a view abandons the traditional aspiration for universal generalizations specifying the necessary and sufficient conditions for an act to be wrong. In its place, it aims to identify feasible ways to bring about incremental progress from our present state. But achieving this aim requires understanding both our present state and what our feasible options are. Kumar argues that it is especially important that we understand facts of the sort that cognitive science is most apt for investigating. He illustrates how to carry out this approach by drawing on the work of John Doris and Peter Railton on character and his own work with Richmond Campbell on "moral consistency reasoning."

Klein agrees with Kumar that cognitive science can contribute to metaethics, debunking arguments, and non-ideal theorizing, but he argues that Kumar's particular approach to non-ideal theorizing presupposes an unduly individualistic model of social psychology: the *Activated Actor* model. In its place, he favors the *Motivated Tactician* model. According to Klein, the crucial difference between these models is how they explain human cognitive limitations. While the Activated Actor model assumes cognitive limitations flow from fixed features of human cognitive architecture, the Motivated Tactician model takes cognitive limitations to be the product of strategic responses to social circumstances that inhibit our ability to achieve our goals.

Because these models take different views on the origins of human limitations, they have different implications for how to overcome limitations in order to promote moral progress. For example, Klein argues that only Motivated Tactician models can explain when and why Kumar and Campbell's moral consistency reasoning promotes moral progress—and when and why it doesn't. Klein argues further that the Motivated Tactician model and the Activated Actor model support different experimental designs and different interpretations of classic findings within moral psychology, which has implications for how to pursue non-ideal moral theorizing. Lastly, he argues that the Motivated Tactician model supports a less individualistic approach to moral development that requires empirical investigation into social facts that go beyond cognitive science's traditional focus on the individual. Klein concludes that, in order for non-ideal theory of the sort envisioned by Kumar to succeed, it must take account of the profound influence of social factors on human moral cognition—and that doing so adequately will require close collaboration between philosophers and cognitive scientists.

★★★

Although the contributions to this volume constitute only a small sample of the work currently being done in philosophy of cognitive science, they reflect some important trends concerning the potential implications of technological innovation (e.g., imaging methods, Bayesian statistical learning models) and increasingly widespread access to existing technologies (e.g., computational power to process large data sets). First, some (e.g., Pullum and Kumar) draw on new empirical methods to shed light on traditional questions by using those methods to support familiar answers to those questions, while others (e.g., Pietroski and Hornstein) do not. Second, some (e.g., Kumar, Klein, Kiverstein, Xu, and Fedyk) draw on new empirical methods to shed light on traditional questions because they think they make salient and support overlooked answers to those questions, while others (e.g., Adams) do not. Third, some (e.g., Smith and colleagues and Kumar) instead use these new empirical methods to upset traditional distinctions, problematize traditional questions that relied on these distinctions, and raise new questions in their wake.

Among those who think new empirical methods can shed light on traditional questions, the question arises: How do the new methods compare to the old ones? For example, Cushman and Niv address this question explicitly when discussing the relative importance of new neuroscientific methods and older behavioral methods, while Kumar and Klein discuss the relative importance of traditional philosophical reflection and new empirical data for adjudicating debates within both philosophy and social psychology.

As new empirical methods become available and philosophers continue to make increasing use of empirical data, controversies in philosophy of cognitive science of the sort discussed in this volume will continue to play an important role in shaping discourse within both philosophy and cognitive science.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part I

Is There a Universal Grammar?



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1 Universal Grammar

Paul Pietroski and Norbert Hornstein

Pity the linguist. By tradition, the job is to describe *languages* in ways that extend the ancient practice of providing grammars (e.g., for Sanskrit or Latin). Modern linguists are also expected to be cognitive scientists who study some aspect of human *psychology* that supports the acquisition and use of languages that children readily attain. Given the tradition, linguists inherit a lot of terminology—e.g., 'sentence', 'subject', 'word', 'language', and 'grammar'—that cannot be discreetly consigned to the attic. So old terms often get used in new ways. As a result, 'Universal Grammar' has meant different things for different people at different times. But the phrase does focus attention on a good idea: The diverse languages that children can naturally acquire exhibit shared features that reflect traits shared by all members of our species.

1. Laying a New Picture on an Old One

Let's start with the obvious fact that nothing does language like humans do language. Other animals communicate. But we talk—a lot. Our distinctive loquacity is manifested along many dimensions, two of which are especially important here. First, humans are linguistically creative, in that we routinely produce and understand expressions that we have never before produced or encountered. Indeed, there seems to be no upper bound on the number of expressions that a native speaker can linguistically manage. Second, the languages that humans can naturally acquire are universally available to children. To be sure, experience matters. Native speakers are only adventitiously proficient in the languages they happen to have acquired. But there are no subgroups of humans who can only acquire languages from certain families (e.g., Indo-European). Given the right experience, any ordinary child can become a speaker of any of the languages that other children can acquire.

These facts are not subtle, and they call for explanation. Moreover, the general form of the required explanation is relatively clear. To account for linguistic creativity, we assume that a native speaker of a language has a grammar for that language. Grammars are *procedures that generate* expressions in certain ways. (We return to some details below.) To account for universal availability, we assume that humans have a mental capacity to acquire grammars of a certain kind, given exposure to speech that is symptomatic of those grammars. This human capacity

is often called the Faculty of Language (FL). The input that FL uses, in producing particular grammars, is often called the primary linguistic data (PLD).

Once these unsubtle facts are noted, it shouldn't be controversial that children acquire grammars by employing a shared FL, which yields various grammars given various courses of PLD. The details are and should be controversial. What kinds of rules characterize the grammars in question? What is the structure of FL? How does FL use PLD to generate any particular grammar? How much of FL is specific to language, and how much is common to other cognitive capacities? These are all legitimate (and hard) questions. But they all presuppose some FL that generates grammars in response to PLD. While it is truistic that FL exists, it takes research to find out what's "in" FL, the PLD, and the grammars that children acquire.

In characterizing the topic this way, we follow Chomsky (1965). On this view, the subject matter of linguistics is doubly psychological: Native speakers have grammars that generate expressions; children acquire these procedures, given their idiosyncratic experience. The goal for linguists is to correctly describe the procedures and the underlying capacity to acquire them. The grammars that a linguist might inscribe on paper are taken to be models of hypothesized expression-generators, and language acquisition is viewed as a process of acquiring an expression-generator given a course of experience and a certain range of options. This suggests a picture in which various grammars can be described with a common vocabulary, making it possible to abstract a common core that can be supplemented in various ways (see Figure 1.1).

This picture presupposes that the child-acquirable languages are importantly alike—despite their manifest diversity—and that many universal features of the grammars that generate these languages reflect species-general features of children. This suggests a second picture, according to which internalized grammars reflect the impact of experience on the cognitive systems that support language acquisition (see Figure 1.2).

Here the idea is that children share a capacity to acquire languages of a certain sort, in response to suitable courses of experience, and acquiring a particular language is a matter of acquiring a mental state that reflects *both* the shared capacity *and* relevant experience.

Chomsky (1965, 1981, 1986) offered proposals that combined the two pictures: Acquiring a particular grammar/procedure is largely a matter of using experience

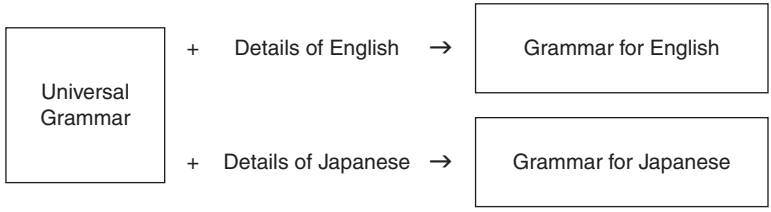


Figure 1.1 Many linguists view language acquisition as a process of acquiring an expression-generator given a course of experience and a certain range of options. Thus, various grammars can be described with a common vocabulary, making it possible to abstract a common core that can be supplemented in various ways.

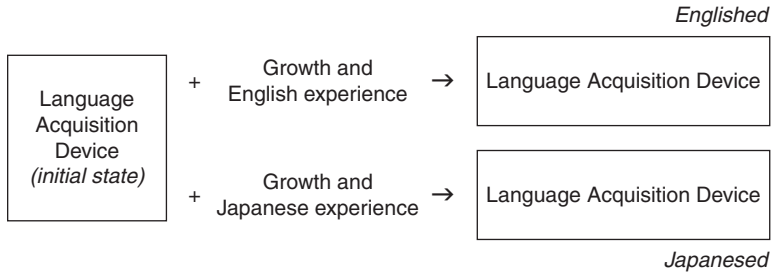


Figure 1.2 Internalized grammars reflect the impact of experience on the cognitive systems that support language acquisition.

to choose from a menu of ranked options, or perhaps fill in a common template for grammars, which exhibit similarities that reflect their shared representational format.[1] Linguists then face the task of capturing the similarities in terms that make it possible to describe the respects in which acquirable grammars vary. This sets the stage for investigating how particular grammars are related to the common core.

2. Unbounded but Constrained

Let's use 'language' generously, allowing for talk of invented mathematical languages and the language of bee dance, along with languages like spoken French and American Sign Language (ASL). We can count something as a language if it connects signals of some kind with interpretations of some kind. Let's use 'Human Language' to talk about the languages that a typical human child can acquire given an ordinary course of growth and experience. These languages connect signals of a special sort—e.g., sounds of spoken French, or gestures of ASL—with interpretations of some kind.

As noted above, Human Languages have some interesting properties that seem to reflect interesting properties of children, who can understand and use strings of words they never previously encountered. In the house that Jack built, there may be a dog that chased a cat that chased a rat, which ate some cheese and thereby led Jack's partner to take drastic action; and so on. Some sentences are too long to say or hear. But children do not acquire languages in a way that imposes an upper bound on how many words can appear in an expression. Counting provides an obvious analogy. The procedure of "adding one" outruns our mortal ability to keep applying it. Similarly, a procedure can connect boundlessly many pronunciations with boundlessly many meanings in a rule-governed way that outruns our limited abilities to process and produce speech. And a child can acquire a *procedure that generates* pronunciation-meaning pairs (π-μ pairs) in an open-ended way.[2]

For many purposes, one can ignore the small differences among the many English procedures, and talk as if nearly a billion individuals share the same

language. The more important point is that if Human Languages are generative procedures, then a proposed grammar "for English" is a hypothesis about how certain internalized procedures generate certain π-μ pairs in an open-ended way, where this way of generating π-μ pairs is one of the ways supported by the Human Language Acquisition Device, depicted above. But even if each Human Language has a lexicon of atomic expressions and at least one combinatorial principle that is somehow recursive, this is not yet a big insight. The interesting questions concern the *kinds* of atomic expressions and *kinds* of combinatorial principles that are permitted—and within those kinds, the range of variation that is possible.

Chomsky (1957) sharpened these questions by describing three kinds of computational systems that can generate boundlessly many expressions, and arguing that Human Languages generate their expressions in some other way; see Lasnik (1999) for a lucid review. This highlighted a pair of fruitful questions: *How* do Human Languages generate π-μ pairs, and *why* do children acquire languages that generate π-μ pairs in those ways, as opposed to other ways?

In addressing these questions, Chomsky drew attention to an important source of data: Speakers know a lot about which meanings they can connect with which pronunciations. For example, string (1) corresponds to a sentence of English, while (2) seems like word salad.

(1) we may have been there
(2) ⋆we have may been there

The unacceptability of (2), indicated with the asterisk, is severe. This suggests that speakers of English *do not* acquire procedures that pair the pronunciation of (2) with a meaning. Contrasts of this sort can provide evidence regarding how Human Languages do and don't generate π-μ pairs. Though as Chomsky (1965) stressed, such unacceptability does not guarantee ungenerability.

Our acceptability judgments may imperfectly reflect what our *languages/grammars* can generate. String (3) also sounds like word salad.

(3) ⋆a rat a cat a dog chased chased squeaked

Nonetheless, speakers of English may well acquire *procedures* that pair the pronunciation of (3) with the meaning of sentence (3a) or some close paraphrase.

(3a) A rat *that was chased by a cat that was chased by a dog* squeaked.

We do acquire procedures that connect the pronunciations of (4) and (5)

(4) a rat a cat chased squeaked
(5) a cat a dog chased chased a rat

with meanings that can be indicated, respectively, with sentences (4a) and (5a).

(4a) A rat *that was chased by a cat* squeaked.
(5a) A cat *that was chased by a dog* chased a rat.

Prima facie, a generative procedure of this sort will also pair the pronunciation of (3) with a meaning like that of (3a).[3] But "center embedding" might impose memory demands that keep us from using our procedures to compute the meaning of (3) in a normal way.[4] In short, (3) may have a meaning that is hard to recognize. In general, you can *have* a procedure that pairs an input with an output, even if circumstances make it hard for you to *use* the procedure as a way of determining the output given the input. It's one thing for a grammar to generate π-μ pairs, in the way that a rule of "adding one" generates number–number pairs, and another thing for a person to connect pronunciations with meanings via spatio-temporally located processes.

That said, (2) doesn't seem to impose any cognitive demand that isn't imposed by (1).

(1) we may have been there
(2) ⋆we have may been there

So while English procedures pair the pronunciation of (1) with the meaning of (1a)

(1a) We might have been there.

they apparently don't pair this meaning—or any meaning—with the pronunciation of (2). Correlatively, string (2) cannot be classified as an English sentence. But the relevant procedures do not merely classify certain word-strings as meaningful sentences. An internalized grammar generates π-μ pairs. So if a proposed grammar is to be a good model of the generative capacity that speakers of English acquire, it must do more than merely identify the word-strings that count as meaningful English sentences. An adequate model will neither *under*generate in the sense of predicting *too few* π-μ pairs, nor *over*generate in the sense of predicting *too many* π-μ pairs.

In this context, consider that Human Languages allow for *homophony*: A single pronunciation can be paired with more than one meaning. For example, the adjective 'bare' shares its pronunciation (***bɛr***) with a noun that can be used to talk about bears and a verb that can be used to talk about bearing weight. More interestingly for purposes of investigating the generative procedures that children acquire, distinct sentences can be homophones, even if they comprise the same string of words. The pronunciation of (6)

(6) the woman saw the man with the telescope

can be understood as having either of the meanings indicated with (6a) and (6b).

(6a) The woman saw the man who had the telescope.
(6b) The woman saw the man by using the telescope.

But note that (6) does not have a third meaning that corresponds to (6c).

(6c) #The woman both saw the man and had the telescope.

In English, 'saw the man with the telescope' *cannot* be understood as a predicate that applies to an individual if and only if 'saw the man' and 'with the telescope' apply to that individual.

We could invent a language in which (6) has all three meanings, or just (6a) and (6c), or just (6a), or just (6c). Correlatively, there are several ways in which a proposed grammar might undergenerate and/or overgenerate meanings with regard to (6). Yet children exposed to ordinary English speech become adults for whom (6) is ambiguous in a *specific limited way*. Perhaps any reasonable procedure compatible with a child's experience will yield a lot of sentential homophony. But then the question is why we didn't acquire grammars according to which (6) has the third meaning. More generally, if the children in a community converge on grammars that don't overgenerate homophony in some respect (compared with other speakers), one wants to know *why* the children acquired grammars that are *restrictive* in this respect.

Consider another example. String (7) has two possible readings; cf. Chomsky (1964).

(7) the goat is ready to eat
  (7a) The goat is prepared to dine.
  (7b) The goat is fit to be consumed.

This contrasts with (8) and (9), where '#' indicates an unavailable interpretation.

(8) the goat is eager to eat
  (8a) The goat is eager to dine.
  (8b) #The goat is eager to be consumed.
(9) the goat is easy to eat
  (9a) #It is easy for the goat to eat.
  (9b) It is easy to eat the goat.

String (8) can only mean that the goat is eager to do some eating. Likewise, (10)

(10) the goat is reluctant to eat

means that the goat is reluctant to do some eating, and not that it is reluctant to be eaten. Regardless of expectations about goats, neither (8) nor (10) is ambiguous in the way that (7) is. String (9) is also unambiguous, but it has an eat-the-goat meaning, as does (11):

(11) the goat is tough to eat

We can invent languages in which (7)–(11) are all ambiguous. So one wants to know why exposure to ordinary English speech *didn't* lead us—or at least many of us—to acquire such languages.

More specifically, there are two hypotheses that theorists need to consider.

(i) Children *could* acquire more permissive grammars, given different experience, but actual experience leads children to acquire suitably restrictive procedures.
(ii) Children *cannot* acquire more permissive grammars, regardless of experience.

Applied to (6), the first hypothesis is that while children can formulate grammars that assign the (c)-meaning, exposure to any ordinary course of English speech will lead to a suitably restrictive grammar. On this view, the Language Acquisition Device (LAD) gives children the resources needed to acquire procedures that overgenerate in this respect, but experience leads children to select more restrictive alternatives. The second hypothesis is that there are relevant constraints on the space of available options: The LAD doesn't let children formulate grammars according to which (6) has the (c)-meaning, so there was no possibility of selecting a procedure that overgenerates in this respect. Similarly, one can hypothesize that children *can but don't* settle on grammars according to which (8) has the (b)-meaning and (9) has the (a)-meaning, or that children *can't and so don't* formulate such grammars. For some cases of "limited homophony," the first hypothesis may be plausible. But if the second hypothesis is more plausible for a wide range of cases, that is an argument in favor of positing a substantive LAD. The conclusion is strengthened if investigation reveals patterns in the limitations.

3. Universals Revealed in Details

Chomsky (1964) discussed illustrative examples like (12),

(12) the woman saw the boy walking towards the house

which can be understood in at least three ways.

(12a) The woman saw the boy while (she was) walking towards the house.
(12b) The woman saw the boy who is walking towards the house.
(12c) The woman saw the boy walk towards the house.

Sentence (12a) implies that the woman walked. Sentences (12b) and (12c) imply that the boy walked. Though unlike (12c), (12b) can be used to describe a situation in which the woman saw the boy without seeing him walk; the boy may have been sitting when the woman saw him. But (13) can only have the interpretation corresponding to (12c).

(13) this is the house that the woman saw the boy walking towards

As a sentence, (13) implies that the woman saw *the boy walk*—not that she saw the boy who is walking (or that she saw the boy while she was walking). The relative clause 'that the woman saw the boy walking towards' is superficially similar to (12). Yet (13) is unambiguous, like (14):

(14) the woman saw the boy walk towards the house
  (14a) #The woman saw the boy and walked towards the house.
  (14b) #The woman saw the boy who walked towards the house.
  (14c) The woman saw the boy do some walking towards the house.

Moreover, (15) is unambiguous in the same way.

(15) what did the woman see the boy walking towards

This string can be used to ask a question corresponding to (12c): What is such that the woman saw *the boy walk* towards it? But (15) cannot be used to ask what is such that she saw the boy who was walking towards it (or such that she saw the boy while she was walking towards it). This suggests a common *constraint* on the meanings of relative clauses and questions, raising the question of why exposure to English speech leads to speakers who assign *only* the see-boy-walk meaning to (13) and (15).[5] If children can formulate more permissive grammars, what stops them from doing so? Absent a plausible answer, the remaining hypothesis is that a child's options are limited, leaving no possibility of selecting a procedure that overgenerates in this respect.

Of course, no single example is decisive. But "negative facts" are ubiquitous. Each string of English words has n meanings, and not $n+1$, for some number n. And as examples like (12)–(15) illustrate, many unattested meanings seem to be constructible from the relevant word meanings in ways that are analogous to actual meanings of superficially similar strings. This is true even for strings that are ungrammatical. For example, (16) is a degraded version of (16a), not (16b); see Higginbotham (1983).

(16) *the child seems sleeping
  (16a) The child seems to be sleeping.
  (16b) #The child seems sleepy.

So if children *could* acquire grammars according to which (16) can be understood/repaired in either way, the question is how ordinary experience consistently leads children to more restrictive grammars. Similar remarks apply to boundlessly many examples, including (13) and (15). When a child acquires a particular language, *many* languages that allow for more homophony go *un*acquired. So however exposure to English speech leads to acquisition of English rather than French, one wants to know how such exposure leads to acquisition of English rather than "English$^+$ languages" that generate more homophony.

Settling on a restrictive grammar, given limited experience, requires some method of "projecting" grammars that generate π-μ pairs in certain restricted ways—as opposed to more permissive grammars that would be equally compatible with (and so not dispreferred by) the limited experience. The data surveyed above suggest that children have a substantive Universal Grammar. Children do not have ready access to data concerning the ways in which strings like (1)–(16)

are or are not ambiguous. Yet children quickly acquire grammars that make these distinctions, while legions of trained linguists—who already know how to talk and ask questions—struggle to formulate grammars that are not woefully inadequate.

This suggests that children have a head start on the linguists, who try to discern the character of this head start by considering data that are unavailable to children. If this is correct, then when a child projects from a particular course of experience to a particular grammar, the child (i) ignores many possibilities that good linguists consider but exclude on the basis of further data, and (ii) considers possibilities that may go unconsidered by generations of linguists. The child enjoys a LAD governed by universal principles, which guide how grammars are projected from the PLD, the restricted linguistic input that the child uses to project a grammar.

Universals, understood in this way, are properties of the LAD that impact the shape of possible generative procedures of human grammars. This technical notion is quite distinct from the notion often employed by linguists who focus on similarities/differences across the surface forms (or construction types) used across Human Languages. There is no reason to think that the two notions should coincide. As Chomsky (1965, p. 118) noted:

> there is no reason to expect uniformity of surface structures, and the findings of modern linguistics are thus not inconsistent with the hypotheses of universal grammarians. Insofar as attention is restricted to surface structures, the most that can be expected is the discovery of statistical tendencies, such as those presented by Greenberg (1963).

Thus, characterizing Universals in Chomsky's sense is not a matter of listing features of Human Languages that are ubiquitous and (easily) observed in the surface forms of every language. The aim is to characterize the principles that delimit a child's options and thereby let children settle on particular grammars by responding to aspects of their experience that *they* attend to, even if the relevant "data" is neither salient to adults nor adequate for purposes of justifying scientific hypotheses about the grammars that children naturally acquire.

4. Gruesome Projections

One can imagine minds that do formulate and acquire grammars by attending to simple examples and projecting rules (e.g., in English, verbs precede objects) that might initially occur to a rational scientist who only had access to the same limited set of data. Such animals would acquire grammars that effectively recapitulate patterns in the data, as opposed to extrapolating in ways that would seem bizarrely unwarranted from the scientist's perspective.

As a way of getting an initial feel for the difference, consider an analogy that Berlinsky (1988) offers, in terms of numbers whose decimal representations are unbounded. Rational numbers, like 10/3 and 22/7, are contrasted with irrational

numbers like π and *e*. The decimal expansions of these numbers are shown below, out to twelve places.

10/3 3.333333333333…
22/7 3.142857142857…
π 3.141592653589…
e 2.718281828459…

The repeating pattern for 10/3 quickly becomes salient. Correlatively, given the twelve-place expansion, reasonable people will agree about the best guess for the thirteenth digit. For 22/7, the pattern emerges more slowly, and a cautious person might want to see the expansion to eighteen places before inferring that the nineteenth digit will be a '1'. For π and *e*, one cannot predict what the next digit will be, unless one knows how to *compute* the decimal expansions. Given the relevant algorithm, the *n*th digit of π is predictable. Likewise, given the relevant algorithm for *e*, one shouldn't expect '1828' to repeat. For these cases, the relevant pattern is specified by an algorithm, but not recapitulated in the series of generated digits. By contrast, for the rational numbers, the pattern determined by the division algorithm is manifested in the series of digits generated by that algorithm; and so for large *n*, the *n*th digit is predictable, given a data set that is limited—say, to expansions of less than twenty places.

In short, a general-purpose "inducer" can make reasonable (or at least non-arbitrary) predictions about "how to go on" given enough initial data concerning 10/3 and 22/7. But with regard to π and *e*, such a system could only guess. So if some system reliably predicts the *n*th digit for the irrational cases, despite a limited data set, then the system is *not* inducing from the data set: It can somehow compute the decimal expansions of certain numbers, and use this capacity to figure out which of these numbers corresponds to the available data. Formulating generalizations that concern chunks of strings will work for some cases, but not all.

Of course, formulating any generalizations requires a suitable vocabulary. And as Goodman (1954) discussed, confirming generalizations by "empirical induction" requires a vocabulary that is suited to this method of *evaluating* generalizations. Initially, one might think that any generalization of the form shown in (17)—e.g., that all emeralds are green—

(17) all Φs are Ψs

is confirmed by a data set that includes many Φs that are also Ψs, and no Φs that are not Ψs (e.g., many green emeralds, and no emeralds that are not green). Any such data set leaves room for the logical possibility that other emeralds are blue, just as '.333333333333' leaves room for the logical possibility that the thirteenth digit is a '7'. But given twelve emeralds, all of which are green, reasonable people will agree about the best guess for the color of the thirteenth emerald. Nonetheless, Goodman showed that not all instances of (17) can be confirmed this way, unless the candidate instances of (17) are restricted to exclude certain logically coherent predicates.

Let *t* be some future moment in time, say the end of today, and define the predicates 'grue' and 'bleen' as follows:

> for any entity e, e is grue if and only if: e is green and observed before *t*, or e is blue and not observed before *t*;
>
> for any entity e, e is bleen if and only if: e is blue and observed before *t*, or e is green and not observed before *t*.

For any emerald that has already been observed, it is grue iff it is green. So if any observed emeralds are uniformly positive instances of (18), they are uniformly positive instances of (19).

(18) all emeralds are green
(19) all emeralds are grue

But while (18) can be confirmed by a data set that includes many emeralds all of which are green, and hence grue, no such data set confirms (19). On the contrary, (19) predicts that many emeralds—i.e., all the emeralds that are still unobserved by the end of today—are blue. And that prediction is surely not confirmed by a data set consisting of uniformly green emeralds.

Vocabulary matters if the goal is to formulate generalizations that can be confirmed by instances of the generalizations. Goodman summarized his point by saying that generalizations like (19) are formulated with *projectable* predicates like 'green', and that the invented predicate 'grue' is not projectable, at least not for creatures who also use predicates like 'blue' and 'emerald.' But it is important to be clear that reference to times is inessential to the issue that Goodman highlighted.

> Let 'G' and 'B' be contrary predicates: $\forall x[G(x) \supset \sim B(x)]$; $\forall x[B(x) \supset \sim G(x)]$. Given a third predicate 'T', we can define a pair of contrary predicates 'G★' and 'B★' as in (20) and (21)

(20) $\forall x\{G^{\star}(x) \equiv [G(x) \,\&\, T(x)] \vee [B(x) \,\&\, \sim T(x)]\}$
(21) $\forall x\{B^{\star}(x) \equiv [B(x) \,\&\, T(x)] \vee [G(x) \,\&\, \sim T(x)]\}$

Alternatively, we can use 'G★' and 'B★' to define 'G' and 'B' as in (20a) and (21a).[6]

(20a) $\forall x\{G(x) \equiv [G^{\star}(x) \,\&\, T(x)] \vee [B^{\star}(x) \,\&\, \sim T(x)]\}$
(21a) $\forall x\{B(x) \equiv [B^{\star}(x) \,\&\, T(x)] \vee [G^{\star}(x) \,\&\, \sim T(x)]\}$

Suppose the domain consists of things on earth, and 'T' applies to things that are north of the equator. Now imagine individuals in the northern hemisphere who endorse generalization (22)

(22) all cows are vegetarians

having noted that every observed cow grazes and never eats meat. If (22) has lots of positive instances, and no negative instances, the "inductive leap" seems reasonable—whether or not you know about hemispheres. Since 'vegetarian' and 'carnivore' are contraries, (22) implies (23).

(23) no cows are carnivores

But if 'vegetarians★' and 'carnivores★' are defined as in (20) and (21), then generalization (24)

(24) all cows are vegetarians★

is unwarranted, despite being like (22) in having each observed cow as a positive instance. For unlike (22), (24) implies that all cows in the southern hemisphere are carnivores, and hence that (23) is *false* if there are any cows in the southern hemisphere.

With regard to 'grue', one can't avoid "cherry picking" the data until time *t*. With regard to 'vegetarian★', one can't avoid cherry picking north of the equator. But the vocabulary-relative character of confirmation remains, even bracketing spatiotemporally indexed predicates. Even if we observe cows from both hemispheres—and, thanks to time travel, even cows observed on the day that the last cows died—there will still be many predicates that all the observed cows satisfy. Let 'T' mark the division between things that moo in a certain way and things that don't. Suppose that all the observed cows happen to moo in this way, which is typical but not exceptionless for cows. A radical skeptic might insist that (22) is unwarranted, given the "biased sample" and the possibility that atypical mooers eat meat. But one can have laxer standards for confirmation without saying that (24) is equally well-confirmed; where now, vegetarians★ are vegetarians that moo in a certain way *or* carnivores that don't moo in that way.

Goodman's "riddle of induction" is to specify how to distinguish generalizations that are confirmed by their positive instances—and the projectible predicates used to formulate such generalizations—from their gruesome counterparts. One can highlight this question by contrasting good generalizations like (22), which seem natural, with crazy generalizations like (24) that we would never think of without help. But one can also view Goodmanian examples as illustrating another point that is directly relevant to the task of describing the space of Human Languages: Confirming generalizations by empirical induction requires a certain kind of vocabulary; and acquiring grammars, in response to ordinary experience, evidently requires a different kind of vocabulary. This somewhat subtle point calls for elaboration.

From a rational scientist's perspective, children make leaps of faith when they settle on grammars that constrain homophony in the specific ways noted in sections two and three. Given a sample of ordinary English speech, who would conclude that (7), (12), and (25) are ambiguous,

(7) the goat is ready to eat
(12) the woman saw the boy walking towards the house
(25) this is the house such that the woman saw the boy walking towards it

but that (8), (13), and (15) are not?

(8) the goat is eager to eat
(13) this is the house that the woman saw the boy walking towards
(15) what did the woman see the boy walking towards

The answer is: any child with human Universal Grammar. Or as Chomsky (1965, p. 24) put it, one who starts a suitable "specification of the class of potential grammars." But it's worth noting how *gruesome* this kind of projection from experience is. As with the decimal expansion of π, the relevant patterns are not discernible in perceptible outputs generated by the relevant procedures.

Consider another analogy. Given two Scrabble tiles in a bag, how often would you need to draw a 'G' (and put it back) before concluding that each of the tiles was a 'G'? Reasonable people can differ. But after seeing a 'G' drawn ten times, who would conclude that the tiles spell 'GO'? The answer is obvious: Someone who is sure that the bag contains an 'O', even if sampling provides no evidence of this. And such a person could reach the same conclusion after seeing a 'G' drawn *once*. From the perspective of someone whose only background assumption is that the bag contains three Scrabble tiles, inferring 'GO' will seem crazy, not merely rash. For someone who knows that the bag contains an 'O', drawing again after seeing a 'G' is pointless; you might as well check to see if any of the local cows are eating blue emeralds. Likewise, children just project grammars as they do, without regard for how reasonable their projections are from a scientific perspective. In short, kids project grammars gruesomely, thereby revealing the character of the Universal Grammar (UG) that both provides and constrains their options.

This raises interesting questions that we cannot address here. In particular, how is the gruesomeness of ordinary language acquisition related to *arithmetic* induction and other historically important examples of evaluating generalizations by non-empirical methods? Students can be led to see that certain generalizations about triangles are theorems that may be *illustrated* with sketchy diagrams that are *not* instances of the theorems. (We also have intuitions regarding what is *possible* but not actual, or *necessary* as opposed to merely actual, despite experience that is limited to the actual.) To what degree are the constraints on Human Languages specific to language, and to what degree are they reflections of uniquely human psychology? These large questions animate investigation. But one can posit a substantive UG while being neutral about how language-specific and human-specific the supported projections are.

5. Bolstering the Point

If all children had access to the same large corpus of π-μ pairs, including relatively rare examples like (15), limitations on homophony would still suggest a substantive UG.

(15) what did the woman see the boy walking towards

Such arguments are reinforced by the fact that children are not constantly scouring the same large data set. Each child has a somewhat idiosyncratic course of experience that unfolds over time, as the child is acquiring one or more grammars, in response to speech that is in many respects less idealized than the set of sentences that appear in a properly edited newspaper. Children have limited memories and attention spans; their capacities for discerning pronunciations and speakers' intentions are not perfect; and so on. This matters. For the relevant question is not how a suitably clever child *could* acquire an English grammar, as opposed to more permissive grammars, given an *ideal* sample of English discourse. Rather, we want to know how a typical child *does* acquire an English grammar given a *typical* sample of English discourse—or more precisely, the temporally unfolding subset of any such sample that corresponds to what a typical child might attend to and represent for purposes of formulating and evaluating grammars.

Moreover, if investigation reveals that young children (say, before the age of 4) have already acquired grammars that generate homophony in constrained ways, then the relevant windows of opportunity for considering and rejecting more permissive grammars are correspondingly compressed. For example, experiments by Crain (2012) and colleagues provide evidence that young children already know a lot about the options for anaphoric dependence of pronouns on candidate antecedents. Examples like (26) and (27)—with coindexing indicating potential dependencies, and asterisks indicating those that are not permissible in English—

(26) Kermit$_1$ said he$_{1/\star 2}$ thinks Grover$_2$ should wash himself$_{\star 1/2}$
(27) Kermit$_1$ said he$_{1/\star 2}$ thinks Grover$_2$ should wash him$_{1/\star 2}$

invite the following generalization: 'himself' *must* have an antecedent that is "nearby," while 'he' and 'him' *cannot* take a nearby antecedent. But 3-year-olds have already settled on a subtler generalization. Like adults, they understand (28) and (29) as unambiguous in the indicated ways.

(28) Kermit$_1$ expected to feed Grover$_2$ and wash himself$_{1/\star 2}$
(29) Kermit$_1$ expected to feed Grover$_2$ and wash him$_{1\star/2}$

If such constraints are manifested early by children, and manifested across superficially different constructions, this makes it even harder to see how typical speech could lead children to reject more permissive grammars. In this regard, consider the contrast between (30) and (31).

(30) who did he say Kermit had criticized
  (30a) For which x: he said Kermit had criticized x.
  (30b) #For which x: x said Kermit had criticized x.
(31) who said he had criticized Grover
  (31a) For which x: x said he had criticized Grover.
  (31b) For which x: x said x had criticized Grover.

As Chomsky (1981) notes, the pronoun 'he' in (30) must be used deictically, to refer to some male individual. But the pronoun can be interpreted as bound in (31); compare (32) and (33).

(32) he said Kermit had criticized everyone
(33) everybody said he had criticized Grover

Data such as these, which indicate that young children are relevantly like adults, suggests a common mechanism for segregating the permissible from illicit interpretations. Whatever this mechanism is, it seems to be operative in young children, at least many of whom have presumably not inspected the relevant linguistic data.

6. Concluding Remarks

Humans are to language like fish are to swimming and birds are to flying. We are built to talk. Relatively simple facts lead quickly to the conclusion that part of this talent is an innate capacity to project grammars (i.e., rules that generate unboundedly many π-μ pairs) from limited examples. In describing the fine structure of this capacity, linguists often use evidence that (i) adult grammars fail to generate certain inductively plausible π-μ pairs, and, in many cases, (ii) children do not try out and then exclude grammars that generate these pairs. Absent pairings often provide valuable clues about the character of relevant mental mechanisms. Correspondingly, theories of UG aim to explain how humans are able to generate the π-μ pairs we do, *and why we don't generate the π-μ pairs we don't*.

In this chapter, we have not addressed questions regarding how linguistically specific UG is. One would like to know how much of the mental apparatus required to acquire a grammar is due to a distinctively linguistic faculty, and how much is due to some combination of more general cognitive capacities. While we bracketed this question here, we think the best way to address it is by discovering principles of UG. Given plausible candidates, one can then ask which aspects of cognition are responsible for these principles. This interesting project is currently part of a research agenda, called the Minimalist Program, which builds on earlier attempts to describe UG as outlined above. Following Chomsky (1995) and others, our hope is that UG will turn out *not* to be very linguistically specific. Saying why is a story for another time. But to answer questions about the sources of UG, one needs an independently plausible conception of UG.

Notes

1 Chomsky (1965, p. 5) quotes Beattie (1788), who said that languages "resemble men" in that though each has peculiarities, whereby it is distinguished from every other, yet all have certain qualities in common. The peculiarities of individual tongues are explained in their respective grammars and dictionaries. Those things that all languages have in common, or that are necessary to every language, are treated of in a science, which some have called *Universal* or *Philosophical* grammar.
2 This leaves room for the possibility that different speakers specify the same π-μ pairs in different ways, much as distinct arithmetic procedures can define the same input-output

pairs. Consider the following variant of "add one": Subtract one; then double the result; then add four; then divide by two. In algebraic terms, $x + 1 = (2(x–1) + 4)/2$. Similarly, $^{+}\sqrt{(x^2–2x + 1)} = |x–1|$. It is a separate question whether *every* Human Language allows for expressions of unbounded length. Embedding might be limited in favor of discourse like 'That dog chased a cat, the cat chased a rat, and the rat ate some cheese.'

3 Similarly, it's hard to see how an English procedure could pair the pronunciation of 'revolutionary new ideas occur infrequently' with a meaning, but not do the same for 'colorless green ideas sleep furiously'. The oddness of the second string presumably reflects our knowledge that its meaning cannot be used to make a sensible claim.

4 Embedding as in (3), where the same type of structure (a reduced relative clause) is repeatedly center embedded, often leads to severe unacceptability.

5 For discussion of perceptual idioms and the relevance of event variables, see Higginbotham (1983). The constraint on extraction from the relative clause 'who was walking towards it' is independently interesting; see Ross (1967).

6 Think of the domain as divided into six regions: (the union of) regions 1, 3, and 5 correspond to (the extension of) 'T'; regions 2, 4, and 6 correspond to '~T'. Let regions 1 and 2 correspond to 'G', while 3 and 4 correspond to 'B'. Regions 5 and 6 correspond to '~G & ~ B'. Then 1 and 4 correspond to 'G★', while 2 and 3 correspond to 'B★'. The extension of 'G★' can be described as a union of intersections: $(\{1, 2\} \cap \{1, 3, 5\}) \cup (\{3, 4\} \cap \{2, 4, 6\})$; i.e., $\{1, 4\}$. But likewise for 'G': $(\{1, 4\} \cap \{1, 3, 5\}) \cup (\{2, 3\} \cap \{2, 4, 6\})$; i.e., $\{1, 2\}$. Similar remarks apply to 'B★' and 'B'. Given the difficulty of learning constraints without "negative data," it is also worth noting that 'grue^{+}' and 'blue^{+}' can be defined so that each of these predicates applies to entity e iff e is blue *or* e is both green and examined before *t*. Then the grue^{+}/blue^{+} things *include* the blue things, as well as any green emerald already examined.

References

Beattie, J. (1788). *Theory of Language*. London.

Berlinksi, D. (1988). *Black Mischief: Language, Life, Logic, Luck*. Boston, MA: Harcourt-Brace.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1964). *Current Issues in Linguistic Theory.* The Hague: Mouton.

Chomsky, N. (1965). Cambridge, MA: MIT Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of Language*. New York: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Crain, S. (2012). *The Emergence of Meaning.* Cambridge: Cambridge University Press.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of Language*, 58–90 (Cambridge, MA: MIT Press).

Higginbotham, J. (1983). The logical form of perceptual reports. *Journal of Philosophy*, 80, 100–27.

Lasnik, H. (1999). *Syntactic Structures Revisited*. Cambridge, MA: MIT Press.

Ross, J. (1967). Constraints on variables in syntax. Doctoral dissertation, Massachusetts Institute of Technology. Published as: Ross, J. R. (1986). *Infinite Syntax!* Norwood, NJ: Ablex.

2 Waiting for Universal Grammar

Geoffrey K. Pullum

Pietroski and Hornstein (this volume; henceforth P&H) see linguists as explorers of the component of the human mind that is responsible for the unfailing success of normal human babies in achieving first language acquisition. They favor a view that is often known as *linguistic nativism*, fashionable among linguists who closely follow the work of Chomsky. Its thesis is that certain innate linguistic prerequisites, possessed by all humans at birth, render language acquisition feasible. P&H group these innate mental characteristics together under the heading of Universal Grammar (UG).[1] But the properties of UG tend to be more boasted of than empirically validated, and P&H supply no new details. This chapter warns readers to heed the warning of Scholz and Pullum (2006) about "irrational nativist exuberance," and draws attention to interesting emergent lines of recent work that P&H do not mention.

Logical Possibilities

P&H follow Chomsky (1965) in regarding a human infant as essentially analogous to a device that, on being exposed to an indefinitely long but finite stream of utterances from some human language, constructs an internal representation of a generative grammar for that language. A generative grammar is a finite system of sentence-building procedures capable of building exactly the sentences of the input language and no others. P&H contend that the task of constructing a generative grammar from the input a child gets would be impossible for an unaided intelligence, but being in possession of the information formalized in the theory of UG makes the task feasible or even straightforward.

I want to concede at the outset that it is certainly possible to imagine a way of responding to a finite input corpus of unprocessed utterances from some language by automatically outputting a correct generative grammar for that language. Imagine a device that internally stores representations of generative grammars for English (G_E), Hawaiian (G_H), and Turkish (G_T), and operates by scanning the acoustic form of input utterances. If the acoustic signature of utterance-final consonants is never encountered, then after a reasonable delay for confirmation it outputs G_H. However, if clear evidence of utterance-final consonants is encountered, G_H is ruled out (since in Hawaiian every syllable ends in a vowel), and thereafter if the characteristic signatures of close front rounded vowels and

close back unrounded vowels are observed with reasonable frequency, it outputs G_T (since Turkish does feature those vowel types). Otherwise, if both of these vowels are lacking, after some reasonable time it outputs G_E.

The device unfailingly produces a correct grammar for the right language, after some exposure to utterances. Yet nothing about grammatical properties has to be learned: Grammars are selected automatically on detection of certain physical properties of acoustic stimuli, and nothing about grammar need be observed at all. (The process could of course be below the level of consciousness: The language acquirer would not need to be aware of anything about its operation.)

It should not be thought that I am inventing a straw man here: The idea that only a finite number of languages need to be considered is not mine. Chomsky (1981, pp. 10–11) proposed very seriously that the learnability of human languages could be guaranteed if only finitely many grammars were allowed by UG, and the idea is referred to as "attractive" by Hornstein (2009, p. 167).[2]

It might also seem strange to depict the infant as never really learning from properties of utterances, but simply jumping involuntarily to certain conclusions under the influence of trigger stimuli. But this too is explicit in defenses of the sort of UG that P&H espouse (Lightfoot, 1989; Gibson and Wexler, 1994; Fodor 1998). Language acquisition is claimed to involve internally scheduled leaps of biological growth, which the environment merely triggers in some cases. Note the remarks of Chomsky (1980, pp. 134–6):

> I would like to suggest that in certain fundamental respects we do not really learn language; rather, grammar grows in the mind … There are certain processes that one thinks of in connection with learning: association, induction, conditioning, hypothesis-formation and confirmation, abstraction and generalization, and so on. It is not clear that these processes play a significant role in the acquisition of language. Therefore, if language is characterized in terms of its distinctive processes, it may well be that language is not learned … It is open to question whether there is much in the natural world that falls under 'learning'.

Logically, it is conceivable that the mental development of both humans and other animals is almost entirely a matter of biologically built-in scheduling, prompted only in some minor respects by sensory experiences, so that "learning" is a folk term with very little applicability. But scientists who believe this need to tell us something about the actual neural architecture of the internally driven growth capacity, and the ways in which experience triggers it. In P&H's chapter we search for that in vain.

What Must Be Learned

One consideration militating against the empirical plausibility of P&H's view is our planet's linguistic diversity, about which they say nothing at all. Human languages turn out to be so diverse in grammatical terms that a tight set of true universal

principles governing them all can hardly be imagined. Some have word formation and inflection processes of extreme complexity: Whole sentences can often be expressed as single words in Eskimoan languages. Others (like Vietnamese) have virtually no word-building. Some (like English) maintain fairly strict constituent order, while others (like Sanskrit, and many aboriginal languages of Australia) have remarkably free word order.

Languages differ, for instance, in the order of Subject (S), Verb (V), and Object (O), in every way they logically could. There are only seven logical possibilities for the normal order for simple, stylistically neutral, declarative clauses, and we find all seven favored in at least some languages: SVO (English, Swahili); SOV (Turkish, Japanese); VSO (Hawaiian, Irish); VOS (Malagasy, Tzotzil); OVS (Hixkaryana, Urarina); OSV (Apurinã, Nadëb), and no strong preference (Sanskrit, Walbiri). Many other syntactic facts also have to be learned without any discernible possibility of significant help from UG: Whether there are prepositions or postpositions (English *in India*, Hindi *Bharat mẽẽ*); modifying adjectives before the noun or after (English *white wine*, French *vin blanc*); determiners before the noun or after (English *the house*, Danish *hus-et*); and so on.

Such differences cannot be brushed aside as minor divergences from a single human language template.[3] Children clearly have to figure out many parochial syntactic facts on the basis of linguistic experience. Since they manage to do it with virtually 100% success, they could surely learn a large array of other facts about normal syntax at the same time, by the same methods of observation, comparison, and familiarization.

Numerous other aspects of a language must clearly be learned from the evidence of experience, since they are so obviously parochial and idiosyncratic. Most obviously, the properties of individual words have to be learned simply by listening to people use them and seeing what happens in the interaction. The learner has to become acquainted with tens of thousands of words, each having properties of many kinds:

- phonology: The plural suffix on *cats* is an entirely different sound from the one on *dogs*; in *insect* the most heavily stressed syllable is the first, but in *infect* it's the second;
- inflection: *Write* has the past participle *written* (not ⋆*writed*); *we* has the accusative form *us* and the genitive form *our*;
- derivational relationships: *Ignorance* denotes the property of being ignorant, but *instance* doesn't denote the property of being instant; *terrified* and *terror* are related in meaning but *rectified* and *rector* are not;
- syntactic properties: *Eat* can have a direct object (*Let's eat it*/*Let's eat*), *devour* must have one (*Let's devour it*/⋆*Let's devour*), *dine* mustn't have one (⋆*Let's dine it*/*Let's dine*); *likely* takes infinitival complements (*He's likely to be late*) but *probable* does not (⋆*He's probable to be late*); *damn* occurs as a modifier before a noun in a noun phrase;
- literal meaning: *Likely* is synonymous with *probable*; *eager* denotes a property that only a mind-possessing entity can exhibit; *damn* adds no truth-conditional meaning;

- conventional implicatures: *Lurking outside* hints at furtiveness or ulterior motive, while *waiting outside* does not; *damn* signals irritation on the utterer's part;
- overtones and associations: *Ain't* is markedly non-standard and colloquial; *fuck* is coarse and offensive; *whilst* is old-fashioned; *whom* is distinctly formal; and so on.

UG cannot help in any substantive way with any of this. There is almost nothing universal about the properties of words: Some of their properties differ dialectally, and even idiolectally (from one speaker to another). For further evidence of the plethora of aspects of human language that cannot plausibly be universalized, see Evans and Levinson (2009)[4] and, with respect to syntax, Culicover (1999, esp. ch. 3).

The Fawlty Strategy

There is a vast literature on approaches to language acquisition with goals other than P&H's; Dąbrowska (2015) offers a very useful survey. But the attitude that P&H seem to maintain toward such alternative literature, and toward research programs that disagree with linguistic nativism, could be called Fawltyism, after the belief of the fictional bigoted British hotelier Basil Fawlty[5] about the key to getting along with Germans: "Don't mention the war!"

One remarkable failure of mention relates to the details of infants' actual linguistic input. P&H point out that we are not interested in "how a suitably clever child *could* acquire an English grammar … given an *ideal* sample of English discourse," but rather, "how a typical child *does* acquire an English grammar given a *typical* sample of English discourse—or more precisely, the temporally unfolding subset of any such sample that corresponds to what a typical child might attend to and represent for purposes of formulating and evaluating grammars." So what sort of sample do children typically get? P&H show absolutely no interest in this question.

Estimates of how much language children hear in their early years run to a million utterances a year or more (Pullum and Scholz, 2002, pp. 44–5, citing Hart and Risley, 1995). And language acquisition begins before a child's first birthday and continues into adolescence (Dąbrowska, 2015, p. 3 and references there). Since some arguments for UG turn crucially on claims about what children almost never encounter (Pullum and Scholz, 2002, pp. 19–23), the quantity and content of this input is highly relevant. Yet P&H ignore it, focusing instead on a small range of invented English utterances illustrating selected syntactic and semantic points.

They argue, for example, that generalizing from *Grover washed himself* (where *himself* must be Grover) and *Grover washed him* (where *him* cannot be Grover) would encourage the child to adopt a false generalization, namely that *himself* must have an antecedent close to it, while *him* is not allowed to. They then show falsity of the generalization with these examples:

Kermit expected to feed Grover and wash himself. (*himself* = Kermit, not Grover)
Kermit expected to feed Grover and wash him. (*him* = Grover, not Kermit)

This would indeed be puzzling if you adopted the practice of looking at nothing more than positions of pronouns and permitted antecedents in word sequences: *Himself* now has its antecedent further away, and *him* refers to a nearby noun phrase—the opposite of what we find in shorter sentences. Yet P&H cite Crain (2012) as having shown experimentally that children as young as 3 can understand such sentences correctly.

But consider how different things are once we assume that children learn not just from overheard word sequences but from the meanings that the context suggests they must have. Reflexive acts like washing oneself are conceptually very different from transitive acts of washing somebody else. Suppose we adopt the neologism *autoablution* for self-washing, to highlight the difference. The first sentence can be paraphrased as "Kermit expected to engage in feeding Grover and autoablution." This does not in any way tempt us to think that washing Grover might be implied: Kermit is the agent, and both the expected activities are his.

P&H's peculiar view is shaped by (i) a refusal to take meaning into account, and (ii) the idiosyncratic fact of English syntax that it expresses reflexive actions like autoablution by using the verb *wash* with a special object pronoun ending in *-self*. Some languages have invariant non-pronouns serving as indicators of reflexivity; many use a form that is also assigned duties like signaling impersonal or a reciprocal meaning (e.g., Spanish *se*); some use affixes on the verb to mark reflexive meanings; some simply mark the verb as detransitivized; some have different constituent orders so that the reflexive element regularly precedes its antecedent; and so on.

What might be universal here is not anything about the syntactic positions of pronouns, but rather the cognitive distinction between reflexive actions and transitive actions—as in the distinction between autoablution and washing an entity distinct from one's own body. Once meanings are considered, not just the linear sequence of words with coreference indices, the different interpretations of *wash himself* and *wash him* become much less puzzling.

Of course, the developmental psychological question of how babies reach the stage of distinguishing utterances about autoablution from those about washing someone still deserves study; babies are known to develop an awareness of agency and transitivity—their own ability to act in a way that affects an object external to themselves—extremely early in life. But it is a characteristic practice of UG defenders to ignore such genuinely psychological facts. P&H prefer to focus attention on bare word sequences (perhaps annotated with coreference indicators), saying little or nothing about interpretation in context, and simply assert that it is unclear how their contrasting properties could be learned.

P&H allude to "a common mechanism for segregating the permissible from illicit interpretations," about which they can say nothing explicit except that: "Whatever this mechanism is, it seems to be operative in young children, at least many of whom have presumably not inspected the relevant linguistic data." It is tautologous to say that there is some kind of mechanism; but the claim about children not inspecting relevant data is impossible to evaluate without intensive study of their actual experience of noticing contextually situated utterances and appreciating the meanings that are being expressed. Most developmental psycholinguists are interested in that kind of study. P&H do not even mention it as a possibility.

Misframed Generalizations

The foregoing example shows us that a puzzle about how some generalization could be learned can depend on the way the generalization is framed. There are indefinitely many ways to describe any given set of facts; we should beware of arguments for nativism that depend on perversely fashioned descriptions. They are common in the nativist literature.

Anderson and Lightfoot (2002, pp. 18–21) provide a particularly clear case, based on facts first noted by King (1970). Several English auxiliary verbs have reduced or 'clitic' forms (phonologically, single consonants; standard spelling writes *'d* for *had* or *would*, *'ll* for *will*, *'m* for *am*, *'re* for *are*, *'s* for *is* or *has*, *'ve* for *have*). These cannot be used everywhere the full forms occur. Anderson and Lightfoot note that maybe you could learn "*is* may be pronounced [z]" just from noting that people sometimes say *Jay is here* and sometimes *Jay's here*, but you would not be able to learn from experience that *I wonder where Jay is* lacks the clitic form **I wonder where Jay's*. In their view:

> The problem is that, because we were not informed about what *cannot* occur, our childhood experience provided no evidence for the "except" clause(s), the cases in which the reduced form is impossible … That is, we had evidence for generalizations like "*is* may be pronounced [z]" … but no evidence for where these generalizations break down.

Their emphasis on our not being "informed about what *cannot* occur" suggests it is impossible to learn that something never occurs without being explicitly told that. I will return to this patently false epistemological claim later. But they also rely on casting the description in a way that makes learnability maximally puzzling, yielding maximal support for a putative innate device of unknown structure and function. Interestingly, though, in this case the description assumed is actually mistaken.

Anderson and Lightfoot attribute the failure of reduced forms to appear in some contexts to the fact that the following piece of the sentence is empty or missing. In *I wonder where Jay is*, the location-denoting phrase that would normally follow *is*, the word *where*, is missing because (as English syntax requires) it is an interrogative word and has to appear at the beginning of its clause. In *Bob isn't interested but Jay is*, a phrase with the meaning "interested" is missing from the second clause. They claim (p. 29) that "where a phrase consists of an auxiliary verb together with a following element such as an adjective, and that following element is unpronounced, the use of a full (rather than clitic) form of the auxiliary is necessary to provide the phrase with enough phonetic substance" to stand alone.

The true generalization, however, has nothing to do with unpronounced following phrases. Zwicky and Pullum (1997) pointed out that the supposed generalization about the blocking of reduced forms is found in certain cases where the complement of the auxiliary is present and pronounced. This is illustrated in the utterances by B below, which are examples of what Zwicky and Pullum call Rejoinder Emphasis:

A: You won't go through with that.
B: I will tóo go through with it! B: *I'll tóo go through with it!
A: She isn't really gonna call the cops.
B: She is tóo gonna call the cops. B: *She's tóo gonna call the cops.

Here the material following *will* and *is* does get pronounced, but still *I'll* and *she's* are not permitted. An alternative analysis is needed to cover these facts. Providing it turns out to make the learnability problem much easier.

The clitic forms of auxiliary verbs, being single consonants, cannot bear even weak stress. But certain constructions require weak stress on an auxiliary verb; for example, Post-Auxiliary Ellipsis, as in *I will* or *She is*, allows a clause to consist of a subject and a weakly stressed auxiliary (and to be understood with an implicit verb phrase meaning following the auxiliary). Rejoinder Emphasis also has such a requirement: It involves a clause consisting of a subject, a weakly stressed auxiliary, a heavily stressed *too* (or sometimes *so*), and (optionally) the appropriate complement for the auxiliary. Because of the weak stress requirement, neither construction can permit clitic auxiliaries.

The learnability problem now melts away: Learners hear *I wíll!*, and *She is tóo gonna*, etc., and use them roughly as they have heard others use them, complete with the weak stress. They never hear sentences like **I'll!* or **She's tóo gonna*, and unsurprisingly they never produce them.

Attempting to draw a moral about UG from facts that have in fact been misdescribed is a common mistake. Take P&H's discussion of *with*-phrases. They assert that a *with*-phrase at the end of a transitive clause may be interpreted as modifying the object (as in *Karen saw her husband with another woman*) or as an instrumental modifier of the verb (as in *Seymour sliced the salami with a knife*). For some sentences either is possible: *The woman saw the man with the telescope* is ambiguous between "saw the man in possession of the telescope" and "saw the man by using the telescope." However, P&H claim, a third meaning, "The woman saw the man and was in possession of the telescope," where the *with*-phrase modifies the subject, is impossible.

They do not state the generalization they are assuming, but it amounts to a claim that a *with*-phrase at the end of a transitive clause cannot be interpreted as applying to the subject. This would entail that *Karen saw her husband with another woman* cannot be made true by a situation where Karen spots her own husband while she is in the company of her friend Maureen. This might seem plausible enough. But the generalization P&H tacitly advocate cannot possibly be imposed by UG; even a tiny extension of the range of data refutes it. Consider just one new case: *Karen watched the eclipse with another woman*. There we do understand the *with*-phrase as modifying the subject of the clause (*Karen*) and not the direct object (*the eclipse*). That is exactly the modification relation that P&H implicitly suggest UG forbids.

What's actually going on has to do with something P&H say nothing about: Contextual plausibility of entailments has a huge influence on what interpretations seem acceptable. Being in the company of a woman is plausible and natural; being in the company of an eclipse is bizarre or impossible. Charity demands that

we select the sensible interpretation. (See Gualmini and Schwarz [2009] for other cases where spurious learnability puzzles are dissolved once pragmatics and charitable interpretation come into the picture.)

Misdirection and Legerdemain

P&H repeatedly point to particular semantic facts about miscellaneous snippets of invented sentences in contemporary standard English, and suggest that it is hard to imagine how infants could learn such facts from experience. They offer not a hint of any general theory of how specific innate cognitive states, mechanisms, or whatever would actually help. Instead they conceal the fact that their explanatory hat contains no rabbit by using smoke, mirrors, and misdirection.

It is certainly true that infants, when exposed to the experience of seeing and hearing utterances in context every day for a year or two, reliably form some generalizations and eschew others. The same is true even when we are older and learn certain less common construction types from reading. When we encounter sentences like (i) and (ii), we form generalizations implying that a sentence like (iii) might also be grammatical:

i. *The more we see her the less we like her.*
ii. *The more they eat the fatter they get.*
iii. *The more it rains the deeper the puddles get.*

But we don't assume that every word in such sentences must contain the letter *e*, or that every such sentence must have an even number of words—both generalizations that happen to hold of (i) and (ii) but not of (iii). The question is not whether indefinitely many useless generalizations go unnoticed when we are forming our general impressions about what we can count on in future experience, but *what are the specific principles or practices that permit us to learn as fast as we do*. That is what we would expect P&H to tell us something about. But they don't.

Indeed, they dismiss or ignore aspects of the acquisition situation that clearly could be of explanatory value. Take the idea that children might tend to produce utterances by combining subparts of utterances they have heard. Far from being carefully rebutted, this is never even mentioned. P&H make the correct observation that *we may have been* is grammatical but **we have may been* is not, as if there were some mystery that babies learn such things. But take any collection of English text you like, and count the occurrences of *may have*, *have been*, *have may*, and *may been*. The first two are extremely common, and the other two basically never occur at all.[6] Thus even if you just made up sentences by randomly stringing together adjacent word pairs that you have definitely heard (which of course is absurdly inadequate as a method for achieving grammatical correctness), you would never construct **we have may been*.

I'm not suggesting that children ever go through a stage of randomly stringing together adjacent word pairs (though in phonology such random trying-out of possibilities does seem to occur; developmental psycholinguists call it the 'babbling' phase). I'm simply pointing to the rather extraordinary fact that P&H

misdirect our attention away from the possibility that what children learn might be influenced by the frequency of word sequences found in what they hear.

P&H also neglect the potential relevance of lexical meaning. They say in footnote 3 that "it's hard to see how an English procedure could pair the pronunciation of 'revolutionary new ideas occur infrequently' with a meaning, but not do the same for 'colorless green ideas sleep furiously'." This does not mention the content of word senses at all. Notice that being revolutionary is compatible with being novel, whereas being colorless is incompatible with being colored green. P&H treat such facts as if they could not possibly be relevant to assigning meanings to word-strings. It seems to me more natural to assume the opposite.

Likewise, when contrasting the ambiguity of *The goat is ready to eat* with the lack of ambiguity in *The goat is eager to eat*, P&H do not mention the fact that readiness is a property of either thinking beings or inanimate objects, while eagerness can only be exhibited by a mind-possessing entity. So your dry-cleaning can be ready for collection but it can't be eager for collection. Thus the relation of readiness can hold between an entity and an eating situation in two ways, but the same is not true for the relation of eagerness.

That doesn't settle the business of how the syntax and semantics of the *ready to eat* construction might be learned (for a terse overview of the rather complex facts, see Huddleston and Pullum 2002, pp. 1256–9). I merely note that P&H never mention that words like *ready* and *eager* (or for that matter *goat* and *eat*) have meanings. Yet no child learns the syntax of infinitival complements of adjectives without paying attention to the meanings of those adjectives. It is reasonable to think that the latter might have some bearing on the former.

Grueness and Induction

Section 4 of P&H's chapter digresses into the philosophy of induction, and the anomalous predicate *grue*, deliberately gerrymandered by Goodman (1954) to be induction-defeating (grueness coincides extensionally with greenness up to a certain arbitrary time point and with blueness thereafter). But all that emerges from this digression is the suggestion that "confirming generalizations by empirical induction requires a certain kind of vocabulary; and acquiring grammars, in response to ordinary experience, evidently requires a different kind of vocabulary."

The reader might expect that some hint of this special vocabulary for grammar induction would follow, but it does not. The hint that induction of grammatical generalizations might succeed because of a constrained vocabulary of induction-friendly predicates is dropped in favor of a different metaphor, one suggestive of either gambling or religion: "Children make leaps of faith when they settle on grammars that constrain homophony in the specific ways noted." This too is promptly dropped in favor of a new analogy about the difference between guessing whether there is an 'O' in a bag of Scrabble tiles and knowing in advance that there is. This leads to a remark about children's grammar acquisition "revealing the character of the Universal Grammar (UG) that both provides and constrains their options."

The metaphors and analogies shift, but it is all metaphors and analogies. P&H give us no glimpse of what UG might actually say. Generative linguists have been talking for half a century (since Chomsky, 1965) about the development of a precise and fully articulated theory of UG, tacit knowledge of which somehow converts an impossible learning task into a feasible triggered acquisition task. P&H do not offer any convincing illustrative example of how this might work.

Computational psycholinguists, meanwhile, are conducting interesting experiments on what is called "unsupervised learning," relying on very few prior assumptions about the structure of languages. There are built-in assumptions, of course—every learning program must incorporate some assumptions about what it has to do—but not up-front assumptions about the structure of sentences or the syntactic and semantic principles shared by all human languages and by no non-human languages, which is what the UG program was always supposed to be about.

Unsupervised learning of human languages proves very hard. Progress has been slow. But the crucial thing is that the rhetorical project of defending UG is not coming up with any components, principles, or insights that help.

Ideas reminiscent of 1940s empiricism turn out to be more useful than anything emerging from UG. Consider an example of a problem so elementary as to be almost trivial by comparison with figuring out the intricacies of syntax and semantics: Identifying the boundaries of words in an utterance assumed to be already analyzed into discrete speech sounds (a generous assumption, since the infant's actual input is a continuous stream of unlabeled noise). The structuralist linguist Zellig Harris (Noam Chomsky's undergraduate and graduate mentor) suggested in 1955 that there was a way to take a string of symbols and find out automatically where the boundaries between meaningful elements are, without appealing to anything about meaning or grammar.

The idea is simple enough. Make a line graph with the symbols of the utterance in temporal sequence along the horizontal axis (without spaces, of course) and positive integers up the vertical axis. Above each transition between segments plot (using statistical data gleaned from observed sequences in other texts) a point indicating how many segments have a fairly high probability of coming next. The graph will have troughs just before any highly predictable symbol (for example, in written English the letter *q* is almost always followed by *u*) and peaks where there are many possibilities for the next symbol (after *ck*, any letter can occur). Harris pointed out that segmenting the utterance at the peaks of the graph gives a remarkably reliable guide to where the word boundaries are (or even the boundaries of the meaningful elements called morphemes, if you target somewhat lower peaks).

In 1955, Harris's idea was just an intuition supported perhaps by a little paper-and-pencil experimentation, but computers are now fast and powerful enough to confirm over large quantities of data that the method works quite well. Harris's intuition has been confirmed several times (see e.g. Tanaka-Ishii and Jin, 2008 and Griffiths et al., 2015). A stream of symbols can be broken into plausible candidates for words or even morphemes on the basis of nothing more than probabilities of specific units at given points in naturally occurring sequences.

Whether or not infants identify the words in their language by computing the positional freedom of occurrence for the phonological units in the sequences they hear is not the issue here (though it should be noted that neonates are in fact sensitive to statistical patterns of frequency in meaningless sound sequences: Saffran et al., 1996). My point is merely that one would expect a progressive UG program to at least have a specific proposal for specific built-in assumptions about identifying wordlike units in utterances, preferably embodying a way in which a learner could identify the meaningful units much faster and with much less work. But that is not happening. UG enthusiasts have proposed no explicit universal word-identification modules to be tested by computational psycholinguists, whereas implementations of Zellig Harris's method have been shown to work.

On the learning of vastly subtler and more complex things in domains like syntax and semantics we are largely ignorant. The mysterious UG module that P&H want us to believe in has not yet contributed anything toward lessening our ignorance.

Abduction and Curiosity

P&H's revisiting of Goodman's "new riddle of induction" yields hardly any payoff. They might have done better to look at some of the modern work on Bayesian abduction—though their Fawltyism biases them against this, since the research program in question is (notoriously) a rival. Bayesian statistical learning is certainly not (in my view) the answer to everything, but I do think that anyone interested in the possibly innate prerequisites of human cognition and language acquisition should be drawing on insights from that literature.

The Bayes–Price rule in probability theory (known as Bayes' Theorem after the Rev. Thomas Bayes, 1701–61) can be interpreted as licensing abductive arguments by relating statistical properties of evidence to plausibility of current beliefs; see chapters 8 and 9 of Nola and Sankey (2007) for an accessible philosophical introduction. How likely is it that a hypothesized grammatical generalization *G* is accurate, given some body of evidence about sentences *E*? Bayes' Theorem entails that the probability of *G* being correct, given *E*, is proportional to the probability that the evidence would look like *E* if *G* were correct. We don't need to get more technical than this to appreciate the crucially important fact about it: It shows that there is a way in which absence of evidence can yield evidence of absence.

It would be fallacious to reason that because you have never observed any instances of some linguistic expression type, it is therefore grammatically impermissible. But it is not fallacious to reason along these lines regarding a type of expression:

i. if expressions like this were allowed, many instances would have turned up by now;
ii. instead there have been none (or negligibly few); therefore (by Bayes' Theorem),
iii. the estimated probability that some grammatical generalization excludes expressions like this should be raised.

Such probabilistic reasoning works even if there are errors in the data. If you have heard thousands of occurrences of *afraid of it* and only very occasional instances of *afraid from it*, that doesn't just confirm the idea that *afraid of it* is allowed by the grammar; it also provides grounds for upping the probability that **afraid from it* is ungrammatical, only occurring sporadically in imperfect English written by foreign speakers.[7] Your statistical experience of not hearing word sequences of certain types provides rational support for the hypothesis that they are grammatically forbidden.

One interesting recent paper by Friston et al. (2017) uses computer-simulation experiments on rule discovery by Bayesian active inference, and relates it to broader aspects of human cognition, notably insight and curiosity. It contains some particularly interesting remarks about curiosity and the drive to acquire knowledge and find order. Knowledge resolves or reduces uncertainty about the world, which is an imperative for any intelligent entity: "any sentient creature must minimize the entropy of its sensory exchanges with the world," where entropy is "uncertainty or expected surprise, where surprise can be expressed as a free energy function of sensations and (Bayesian) beliefs about their causes." And "resolving different sorts of uncertainty furnishes principled explanations for different sorts of behavior."

Through Bayesian model reduction Friston and colleagues then show, in effect, how learning is enormously facilitated by the assumption that *there are rules or symmetries*—the assumption that the universe is not random. This might be seen as echoing the final remarks of Herbert Feigl in his classic 1934 paper on the logic of the principle of induction:

> The attempt to know, to grasp an order, to adjust ourselves to the world in which we are embedded, is just as genuine as, indeed, is identical with, the attempt to live. Confronted with a totally different universe, we would nonetheless try again and again to generalize from the known to the unknown. Only if extended and strenuous efforts led invariably to complete failure, would we abandon the hope of finding order. And even that would be an induction.

Friston et al. and Feigl seem to see human beings, even more than any other animals, as constantly (albeit subconsciously) striving to reduce uncertainty not just by gathering evidence about the world of their experience but also by actively developing a formulation of it that permits explanation of what occurs and why. This applies to linguistic experience just as much as to other kinds of experience: We search for an explanatory understanding of the verbal behavior that we observe, one that permits us to see why other people's utterances have the form they do—and don't have the form they don't.

P&H's claim is that data-driven hypothesizing could not possibly work for language: Some hypotheses must be ruled out *a priori* for humans, not on grounds of not fitting the facts but because UG simply won't permit them. An imperative on the part of all sentient beings to seek regularities in experience would not be enough, in P&H's view, to account for what happens: Human

languages must be designed in a certain special way that would be undiscoverable without a special component of mind supplying crucial clues about what generalizations to adopt.

That view overlooks another whole research program that P&H's Fawltyism precludes them from citing. The work of Simon Kirby and his collaborators over the past two decades has suggested that we should "concentrate less on the way in which we as a species have adapted to the task of using language and more on the ways in which languages adapt to being better passed on by us" (Kirby, 2001, p. 110). In a long series of ingenious experiments on human subjects' ability to learn invented languages (see Kirby, 2017, for a recent overview), Kirby and colleagues have developed a view very different from the UG one. For example, when human subjects attempt to learn a mini-language with random word/meaning associations, and the errors made by each 'generation' of learners are built into the version of the language handed to the next batch of subjects, regularities in the relation between word structure and meaning components begin to emerge over the 'generations.' The UG mindset sees our progress as assisted by our being restricted to a class of languages that UG deems fit and proper for us. The findings of Kirby and colleagues suggest rather that our cognitive shortcomings (like our limited memory for random irregularity) slowly sculpt languages to make them more learnable by creatures like us.

Of course it is still logically possible that some of what we have yet to discover about language acquisition will turn out to involve inbuilt features of human brains specifically tied to language and nothing else. However, imagining that such features have already been discovered, and that linguistic research has mapped their structure and set out a theory of UG that reveals how they function to aid the acquisition process, would be a major misunderstanding of the state of the art in the cognitive and linguistic sciences.

Acknowledgment

I am grateful to Bob Borsley, Jenny Culbertson, Jim Donaldson, Karl Friston, and Mark Steedman for very useful critical comments. They agree neither with each other nor with me, so what I say here must be regarded as purely my responsibility, not theirs.

Notes

1 P&H introduce the term 'Universal Grammar' in the first section of their chapter, apologizing for its ambiguity, but then talk about a 'Faculty of Language' (FL) and a Language Acquisition Device (LAD), returning to introduce the abbreviation 'UG' only near the end. If I understand their intent correctly, UG is the theory of what is in the FL and thus constrains the LAD.

2 See Pullum (1983) for a detailed argument against trying to define UG in a way that limits the class of grammars (hence languages) to a finite set. The suggestions for how a finite bound might be achieved would certainly allow for an astronomically huge number of distinct grammars, too large for finiteness to be of any use. The finitely-many-languages idea is seldom mentioned in the contemporary literature.

3 Chomsky remarks that "even down to fine detail, languages are cast to the same mold" and an unbiased scientist from Mars "might reasonably conclude that there is a single human language, with differences only at the margins" (2000, p. 7). Current knowledge about the remarkable typological diversity of human languages makes that look extremely implausible to me, for a Martian investigator even minimally attentive to word and sentence structure.
4 Evans and Levinson's title ("The myth of language universals") is ill-chosen: Their central point is that the sheer diversity of human languages may be more interesting for cognitive scientists than whatever properties languages turn out to share.
5 From the 1975 BBC TV situation comedy *Fawlty Towers*, series 1, episode 6: "The Germans."
6 I did a quick check on 44 million words of newspaper text and about a million words of classic novels. The frequency of *have may* sequences is literally zero. The sequence can occur accidentally in sentences of more than one clause, of course (*What they have may not suit you*), but such sentences are in practice amazingly rare.
7 Google reports some 50 million web hits for *afraid of it*, but only 35,000 for *afraid from it* (and, for the web, that is close to zero). Most of the latter are mistakes; one token reads: "All of us thinking about death maybe because we afraid from it or because we thinking that our death more good than our current life."

References

Anderson, S., and Lightfoot, D. W. (2002). *The Language Organ: Linguistics as Cognitive Physiology*. Cambridge: Cambridge University Press.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.
Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.
Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
Crain, S. (2012). *The Emergence of Meaning.* Cambridge: Cambridge University Press.
Culicover, P. W. (1999). *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquisition*. Oxford: Oxford University Press.
Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, 6, article 852.
Evans, N. and Stephen Levinson, S. (2009). The myth of language universals. *Behavioral and Brain Sciences*, 32, 429–92. doi: 10.107/S0140525X0999094X
Feigl, H. (1934). The logical character of the principle of induction. *Philosophy of Science*, 1, 20–9. doi: 10.1086/286303. Reprinted in Herbert Feigl and Wilfrid Sellars (Eds.), *Readings in Philosophical Analysis*, 297–304 (New York: Appleton-Century-Crofts, 1949).
Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, 29, 1–36.
Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity, and insight. *Neural Computation*, 29, 1–29.
Gibson, E., and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25, 407–54.
Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.
Griffiths, S., Purver, M., and Wiggins, G. A. (2015). From phoneme to morpheme: A computational model. In H. Baayen, G. Jäger, M. Köllner, J. Wahle, and A. Baayen-Oudshoorn (Eds.), *6th Quantitative Investigations in Theoretical Linguistics Conference*. Tübingen, Germany. http://doi.org/10.15496/publikation-8639
Gualmini, A., and Schwarz, B. (2009). Solving learnability problems in the acquisition of semantics. *Journal of Semantics*, 26, 185–215.

Harris, Z. (1955). From phoneme to morpheme. *Language*, 31, 190–222.

Hart, B., and Risley, T. (1995). *Meaningful Differences in the Everyday Experiences of Young Children*. Baltimore, MD: Paul H. Brookes.

Hornstein, N. (2009). *A Theory of Syntax*. Cambridge: Cambridge University Press.

Huddleston, R., and Pullum, G. K. et al. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

King, H. V. (1970). On blocking the rules for contraction in English. *Linguistic Inquiry*, 1: 134–6.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5, 102–10.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin and Review*, 24, 118–37.

Lightfoot, D. (1989). The child's trigger experience: Degree-0 learnability. *Behavioral and Brain Sciences*, 12, 321–34.

Nola, R., and Sankey, H. (2007). *Theories of Scientific Method: An Introduction*. Montreal: McGill-Queen's University Press.

Pullum, G. K. (1983). How many possible human languages are there? *Linguistic Inquiry*, 14, 447–67.

Pullum, G. K., and Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review*, 19, 9–50.

Pullum, G. K., and Zwicky, A. M. (1997). Licensing of prosodic features by syntactic rules: The key to auxiliary reduction. Paper presented at the annual meeting of the Linguistic Society of America. http://www.lel.ed.ac.uk/~gpullum/PZ1997.pdf

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274 (13 December), 1926–8.

Scholz, B. C., and Pullum, G. K. (2006). Irrational nativist exuberance. In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science*, 59–80. Oxford: Basil Blackwell.

Tanaka-Ishii, K., and Jin, Z. (2008) From phoneme to morpheme—another verification in English and Chinese using corpora. *Studia Linguistica*, 62, 224–48. doi: 10.1111/j.1467-9582.2007.00138.x

Further Readings for Part I

Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2014). Child language acquisition: Why universal grammar doesn't help. *Language*, 90(3), e53–e90. https://doi.org/10.1353/lan.2014.0051

Argues that no extant accounts of Universal Grammar succeed in simplifying learning various aspects of language on the basis of limited linguistic input.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, 35(7), 1207–42. https://doi.org/10.1111/j.1551-6709.2011.01189.x

Defends the existence of Universal Grammar by arguing against recent attempts to show that domain-general learning mechanisms could explain language acquisition on the basis of limited linguistic input.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

The seminal work in which Chomsky introduces and argues for the existence of Universal Grammar.

Crain, S., & Pietroski, P. (2001). Nature, nurture and Universal Grammar. *Linguistics and Philosophy*, 24(2), 139–86. https://doi.org/10.1023/A:1005694100138

Buttresses classic poverty-of-the-stimulus arguments for Universal Grammar by relying on recent work in theoretical linguistics and psycholinguistic studies of child language.

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anybody seen it? *Frontiers in Psychology*, 6, 852.

Argues that the most prominent arguments for the existence of Universal Grammar are based on assumptions that are false or unsupported by the available evidence.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–48. https://doi.org/10.1017/S0140525X0999094X

Reviews a large body of work in comparative linguistics revealing how languages differ from one another at various levels and along various dimensions, and uses this to argue that Universal Grammar hypotheses are refuted by empirical data, unfalsifiable, or else nearly vacuous. Includes a number of insightful peer commentaries and replies from the authors of the target article.

Everett, D. L. (2012). What does Pirahã grammar have to teach us about human language and the mind? *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(6), 555–63. https://doi.org/10.1002/wcs.1195

A review of studies on the Pirahã language that argues that recursion is not necessary for human language and argues against the existence of Universal Grammar.

Fodor, J. D., & Sakas, W. (2016). Learnability. In I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar*, 249–69. Oxford: Oxford University Press.

Compares and contrasts various models of how children use experience to fix specific parameter values and acquire a grammar from the options made available by Universal Grammar.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour*, 1(10), 713. https://doi.org/10.1038/s41562-017-0184-4

Reviews brain imaging studies on the neural basis of language and argues that the imaging data is consistent with a view of the linguistic faculty that posits Universal Grammar.

Goldberg, A., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4), 468–77. https://doi.org/10.1002/wcs.22

Reviews usage-based approaches to language acquisition, according to which domain-general capacities can explain both diversity and commonality across languages without the aid of Universal Grammar.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–79. https://doi.org/10.1126/science.298.5598.1569

Relies on work in evolutionary biology, anthropology, psychology, and neuroscience to argue that it is only the capacity for recursive computation that distinguishes the human faculty of language from animal communication systems, and it explores possible evolutionary origins of this capacity.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10), 451–6. https://doi.org/10.1016/j.tics.2004.08.006

Synthesizes arguments for and against Universal Grammar to argue that language acquisition requires statistical learning, but that it also requires Universal Grammar to specify syntactic parameters along which learning takes place.

Study Questions for Part I

1) How do Pietroski and Hornstein characterize Universal Grammar?
2) According to Pietroski and Hornstein, what is universal availability and linguistic creativity? Why do they think these phenomena support the existence of Universal Grammar?
3) What kinds of constraints on linguistic creativity do Pietroski and Hornstein invoke to argue for the existence of Universal Grammar? Why do they think that the existence of these constraints supports the existence of Universal Grammar?
4) How does the data that Pietroski and Hornstein draw on to argue for the existence of Universal Grammar differ from the data that Pullum draws on to argue against the existence of Universal Grammar?
5) According to Pullum, why does the data he reviews undermine the existence of Universal Grammar?
6) According to Pietroski and Hornstein, why does positing Universal Grammar help explain language acquisition? According to Pullum, why does positing Universal Grammar fail to explain language acquisition?

Part II

Are All Concepts Learned?



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3 Beyond Origins

Developmental Pathways and the Dynamics of Brain Networks

Linda B. Smith, Lisa Byrge, and Olaf Sporns

The invitation asked that we write on the question of whether concepts are innate or learned, with the idea being that we would take the side that they are learned. We cannot do the assigned task because the question, while common, is ill-posed. The terms *concept* and *innate* lack formal definitions. *Innate* has no standing within 21st century biological developmental processes (Stiles, 2008). *Learning*, in contrast, has many formally specified definitions (e.g., habituation, reinforcement learning, supervised learning, unsupervised learning, deep learning, and so forth; see Hinton, Osindero, & The, 2006; Michalski, Carbonell, & Mitchell, 2013) that have found many real-world applications in machine learning and have analogues in some forms of human learning. But surely the intended question was not whether the human mind is equivalent to implementations of machine-learning algorithms or in some way different. If so, our answer would be "different." Instead, we interpret the intended question to be about the *development* of the basic architecture of human cognition.

In brief, our answer to that question is this: Human cognition emerges from complex patterns of neural activity that, in fundamentally important ways, depend on an individual's developmental and experiential history (Byrge, Sporns, & Smith, 2014). Dynamic neural activity, in turn, unfolds within distributed structural and functional brain networks (Byrge et al., 2014). Theory and data implicate changing brain connectivity (i.e., changing brain networks) as *both* cause and consequence of the developmental changes evident in human cognition. These developmental changes emerge within a larger dynamic context, constituting a brain–body–environment network, and this larger network also changes across time. Understanding the development of human cognition thus requires an understanding of how brain networks together with dynamically interwoven processes that extend from the brain through the body into the world shape developmental pathways (Thelen & Smith, 1994; Chiel & Beer, 1997; Byrge et al., 2014).

In the chapter that follows, we attempt to flesh out this contemporary understanding of the "origins" of human cognition. We conclude that the traditional notions of "concepts" and "innateness" have no role to play in the study of the mind.

Brain Networks

There are specialized brain regions that have been associated with specific cognitive competencies; however, research over the last 20 years has shown that different brain regions cooperate with one another to yield systematic patterns of co-activation in different cognitive tasks (Sporns, 2011). These patterns of cooperation depend on and reveal two kinds of brain networks: structural and functional networks. Structural networks are constituted by anatomical connections linking distinct cortical and subcortical brain regions. Functional networks are constituted by statistical dependencies among temporal patterns of neural activity that emerge in tasks but are also evident in task-free contexts (also called resting-state connectivity). For example, during reading, when left inferior occipitotemporal regions (linked with visual letter recognition) are active, temporally correlated evoked activity is also observed in left posterior superior temporal cortex (linked with comprehension) and in left inferior frontal gyrus (linked with pronunciation, see Dehaene et al., 2010). These regions thus form part of a "reading functional network" and jointly coordinate their activity during reading. Parts of this reading network are also involved in other functional networks, including spoken language production and on-line sentence processing (Dehaene et al., 2010; Monzalvo & Dehaene-Lambertz, 2013). This is the general pattern for all of human cognition; all of our various everyday activities recruit different assemblies of neural components, and so each individual component is involved in many different kinds of tasks.

Over time, brain networks change in some respects and remain stable in others. Structural networks are relatively stable, but they do change over the longer time scales of days, weeks, and years. Functional networks are much more variable, especially when observed on short time scales, but they also exhibit highly reproducible features over time, as can be seen in resting-state connectivity patterns. Changes in functional networks can therefore be measured over multiple time scales. At the time scale of milliseconds to seconds, functional networks undergo continual change, reflecting spontaneous and task-evoked fluctuations of neural activity (see Byrge et al., 2014 for a review). Over longer time intervals of several minutes, functional networks exhibit robustly stable features across and within individuals even at rest, features that are thought to reflect the brain's intrinsic functional architecture (Raichle, 2010). Nonetheless, these more stable features of functional networks also change over longer time scales, in response to the history of individuals (e.g., Tambini, Ketz & Davichi, 2010; Harmelech, Preminger, Wertman, & Malach, 2013; Mackey, Singley, & Bunge, 2013).

There are many examples of such long-term change. For example, perceptual learning changes the psychophysics of discrimination by changing the degree to which spontaneous activity between networks is correlated (e.g., Lewis et al., 2009). Likewise, mastering challenging motor tasks or musical training has been shown to lead to long-lasting changes in functional resting-state networks (Dayan & Cohen, 2011; Luo et al., 2012). Finally, when children learn to recognize and write letters, this leads to changes in the co-activation of motor and visual areas

(James & Engelhardt, 2012). Importantly, these changes in functional networks do not just affect one task but have cascading consequences for many tasks. This is because stimulus-evoked activity is always a perturbation of ongoing activity. Therefore, experience-dependent changes in patterns of ongoing (resting-state) functional connectivity in the brain may have an effect on the potential response of the system to intrinsic and extrinsic inputs.

Structural and functional networks also interact. On short time scales, structural and functional networks mutually shape and constrain one another within the brain. On long time scales, both generate and are modulated by patterns of behavior and learning. But, over the longer term, and as a consequence of their own activity in tasks, these networks change. For instance, moment-to-moment fluctuations in intrinsic functional connectivity predict moment-to-moment variations in performance on tasks such as ambiguous perceptual decisions and detection of stimuli at threshold (see Fox & Raichle, 2007; Sadaghiani, Poline, Kleinschmidt, & D'Esposito, 2015). Further, there have been many demonstrations of experience-induced changes in brain networks, and these studies reveal that individual differences in both structural and functional brain networks are associated with differences in cognitive and behavioral performance (see Deco, Tononi, Boly & Kringelbach, 2015). Our purpose in reviewing these findings is to show that understanding human cognition—its potential and its constraints—requires understanding the multi-scale dynamics of brain networks.

The dynamic properties of brain networks also clarify conceptual issues relevant to the "origins" of human cognition. First, the role of connectivity goes beyond channeling specific information between functionally specialized brain regions. In addition, connectivity generates complex system-wide dynamics that enable local regions to participate across a broad range of tasks. Second, the role of external inputs goes beyond the triggering or activating of specific subroutines of neural processing that are encapsulated in local regions. Inputs act as perturbations of ongoing activity whose widespread effects depend on how these inputs become integrated with the system's current dynamic state (Fontanini & Katz, 2008; Destexhe, 2011). Third, the role of connectivity and the role of experience go beyond enabling the performance of specific tasks. They also produce alterations in the spontaneous (resting-state) activity across these networks, and these alterations can influence the response of the system in novel tasks (Byrge et al., 2014). Fourth, the cumulative history of perturbations as recorded in changing patterns of connectivity—in-the-moment and over progressively longer time scales—defines the system's changing capacity both to respond to input and to generate increasingly rich internal dynamics.

A long time ago, with no knowledge of the dynamics of the human brain, William James (1890/1950) had it fundamentally right when he wrote:

> our brain changes, and … like aurora borealis, its whole internal equilibrium shifts with every pulse of change. The precise nature of the shifting at a given moment is a product of many factors … But just as one of them is certainly the influence of the outward objects on the sense-organs during the moment,

> so is another certainly the very special susceptibility in which the organ has been left at that moment by all it has gone through in the past.
>
> (James, 1950, p. 234)

Brain–Body–Environment

The changes in the brain—over the short term of momentary responses and over the longer term of developmental process—cannot be understood by studying the brain in isolation. Brain networks do not arise autonomously, but rather as the product of intrinsic and evoked dynamics, local and global neural processing, through constant interaction between brain, body and environment.

First, brain networks drive real-time behavior; this behavior in turn evokes neural activity that can change patterns of connectivity. For instance, when we hold a cup, write our name, or read a book, different but overlapping sets of neural regions become functionally connected and co-activation patterns emerge across participating components (Sporns, 2011). At multiple time scales, these co-activation patterns evoke changes within components and across the networks that extend beyond the moment of co-activation. This leads to further enduring functional and structural changes. Evoked neural activity from performing even relatively brief tasks such as looking at images causes perturbations to intrinsic activity that last from minutes to hours (e.g., Han et al., 2008; Northoff, Qin, & Nakao, 2010) and are functionally relevant, predicting later memory for the seen images (Tambini et al., 2010). These "reverberations" of evoked activity may also modulate structural topology via longer-lasting changes in synaptic plasticity and thus downstream activation patterns. Extensive practice in tasks such as juggling produces changes in the structure of cerebral white matter (Sampaio-Baptista et al., 2013) over slow time scales of weeks and longer (Zatorre, Fields, & Johansen-Berg, 2012) with task-induced modulations of functional and structural connectivity occurring in tandem (Taubert et al., 2011). All of this strongly suggests that an individual brain's network topology and dynamics at one time point reflect a cumulative history of its own past activity in generating behavior.

Second, behavior does not stop at the body but also physically affects and makes changes in the world (Clark, 1998). Behavior extends brain networks into the environment, coupling brain activity in real time to sensory inputs. Coordinated, distributed neural activity generates behavior. By its perceptible effects on the world—an object moved, a noise heard, a smile elicited—that behavior evokes coordinated neural activity across the brain. This real-time interaction of brain, behavior, and sensory inputs dynamically couples different regions in the neural system, modulates functional connectivity, and thereby drives change across functional and structural networks. We use Figure 3.1 to illustrate these ideas. When we hold and look at an object, brain networks drive the coordination of hand, head, and eye movements to the held object. As we interact with that object, through moving eyes, heads, and hands, we actively generate dynamic sensory-motor information that drives and perturbs neural activity and patterns of connectivity.

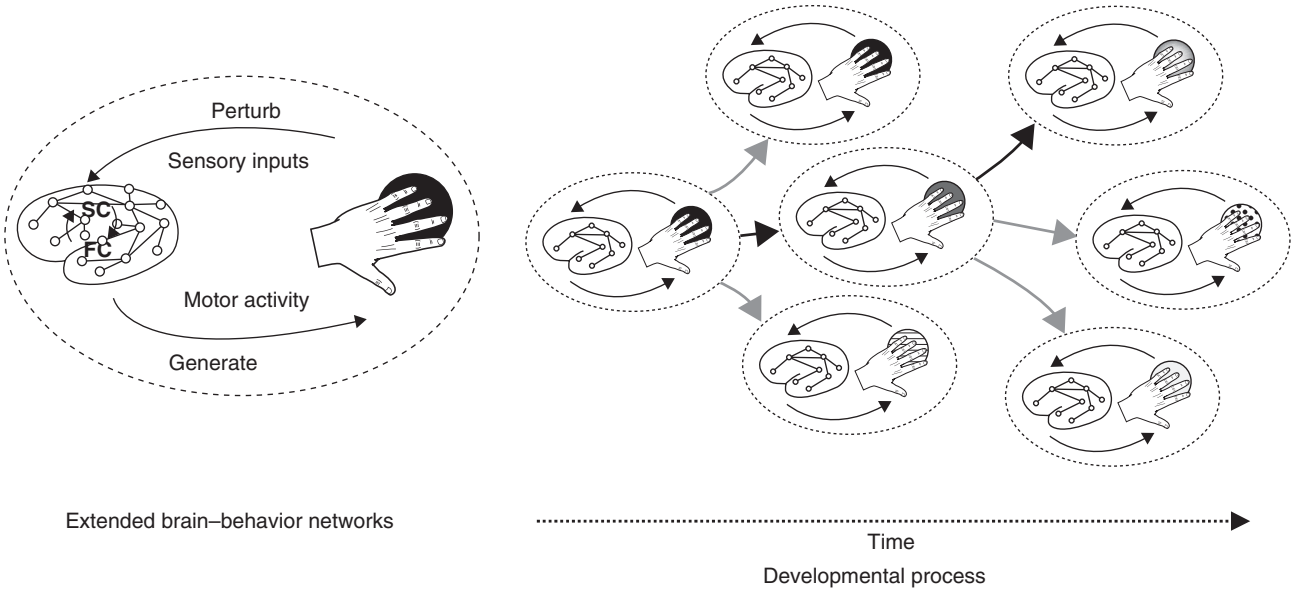


Figure 3.1 Extended brain–body–behavior networks mutually shape and constrain one another across time scales, with the developmental process emerging from these multi-scale interactions. Left: Behavior extends brain networks into the world by selecting inputs that perturb the interplay between structural (SC) and functional (FC) networks within the brain. These stimulus-evoked perturbations cascade into intrinsic brain dynamics, producing changes in functional and structural networks over short and long time scales, changes that modulate subsequent behavior. Right: These extended brain–behavior networks undergo profound changes over development, with changes in the dynamics of the body and behavior (e.g. sitting, crawling, walking, or reading) creating different regularities in the input to the brain—and in turn modulating functional and structural networks of the brain, which in turn modify later behavioral patterns.

There is a remarkable amount of behavioral data from developmental psychology consistent with these claims (illustrated in Figure 3.1) about object manipulation, in particular. In human infants, recognition of the three-dimensional structure of object shape depends on and is built from the rotational information generated by the infant's own object manipulations (e.g., Pereira, James, Jones, & Smith, 2010; Soska, Adolph, & Johnson, 2010; James, Jones, Smith, & Swain, 2014). But the core developmental idea is older. Piaget (1952) described a pattern of infant activity that he called a secondary circular reaction. A rattle would be placed in a 4-month-old infant's hands. As the infant moved the rattle, it would both come into sight and make a noise. This would arouse and agitate the infant, causing more body motions, which would in turn cause the rattle to move into and out of sight and to make more noise. Infants at this age have very little organized control over hand and eye movement. They cannot yet reach for a rattle and, if given one, they do not necessarily shake it. But if the infant accidentally moves it, and sees and hears the consequences, the activity will capture the infant's attention—moving and shaking, looking and listening—and through this repeated action the infant will incrementally gain intentional control over the shaking of the rattle.

Piaget thought this pattern of activity—an accidental action that leads to an interesting and arousing outcome and thus more activity and repeated experience of the outcome—to be foundational to development itself. Circular reactions are perception–action loops that create opportunities for learning. In the case of the rattle, the repeated activity teaches the infant how to control their body, which actions bring held objects into view, and how sights, sounds, and actions correspond. Piaget believed this pattern of activity, involving multimodal perception–action loops, to hold the key to understanding the origins of human cognition. The core idea of a cyclical reaction and its driving force on development is now understandable in the dynamics of brain networks. Holding and shaking the rattle couples different brain regions, creating a network, both in the generation of that behavior as well as in the dynamically linked sensory inputs created by its effects upon the world.

This is an example of the fundamental role played by a brain–body–environment network and it is the answer, just as Piaget saw, to the "origins" question (Sheya & Smith, 2010). Sampling of the external world through action creates structure in the input, which in turn perturbs ongoing brain activity, modulating future behavior and input statistics, and changing both structural and functional connectivity patterns. But this active sampling of the world is itself driven by neural activity, as motor neurons modulated by intrinsic neural activity and network topology guide the movements of eyes, heads, and hands. The brain's outputs influence its inputs, and these inputs in turn shape subsequent outputs, binding brain networks—through the body—to the environment over short time scales, and cumulatively over the course of development.

Developing Brain–Body–Behavior Networks

With development, changes are seen across all aspects of this cyclical process: the brain, its outputs, and its inputs. The development of brain networks is

protracted, extending from postnatal pruning and myelination to synaptic tuning and remodeling over the lifespan (Stiles, 2008; Hagman, Grant, & Fair, 2012). In early human development, the body's morphology and behavior change concurrently, which results in continual but developmentally ordered changes in the input statistics. Figure 3.2 illustrates the dramatic changes in the motor abilities of humans over the first 18 months of life. A large literature documents dependencies between these specific motor achievements and changes in perception and other developments in typically (see Bertenthal & Campos, 1990; Smith, 2013; Adolph & Robinson, 2015) and atypically developing children (Bhat, Landa, & Galloway, 2011). For example, pre-crawlers, crawlers, and walkers have different experiences with objects, different visual spatial experiences, different social experiences, and different language experiences that are tied to posture and can be influenced by experimentally changing the infant's posture (Adolph et al, 2008; Smith, Yu, Yoshida, & Fausey, 2015,). Input statistics change profoundly with every change in motor development, and the constraints of the developing body on the brain–body–environment network may be essential to explaining why human cognition has the properties that it does.

Recent research in egocentric vision provides a strong case study. Egocentric vision is the first-person view, as illustrated in Figure 3.2. The first-person view has unique properties and is highly selective because it depends on the individual's momentary location, orientation in space, and posture (see Smith, Yu, Yoshida, & Fausey, 2015, for review). First, the scenes directly in front of infants are highly selective with respect to the visual information in the larger environment (e.g., Yu & Smith, 2012; Smith et al., 2015). Second, the properties of these scenes differ systematically from both adult-perspective scenes (e.g., Smith, Yu, & Pereira, 2011) and third-person perspective scenes (e.g., Yoshida & Smith, 2008; Aslin, 2009; Yurovsky, Smith, & Yu, 2013), and they are not easily predicted by adult intuitions (e.g., Franchak, Kretch, Soska, & Adolph, 2011; Yurovsky et al., 2013). Third, and most critically, the information and regularities in these scenes are different for children of different ages and developmental abilities (Kretch, Franchak, & Adolph, 2014; Jayaraman, Fausey, & Smith, 2015; Fausey, Jayaraman, & Smith, 2016).

Infant-perspective scenes change systematically with development because they depend on the perceiver's body morphology, typical postures and motor skills, abilities, interests, motivations, and caretaking needs. These all change dramatically over the first two years of life, and thus collectively serve as developmental gates to different kinds of visual data sets. In this way, sensory-motor development bundles visual experiences into separate data sets for infant learners. For example, people are persistently in the near vicinity of infants during their first two years and people have both faces and hands connected to the same body. But analyses of a large corpus (Fausey et al., 2016) of infant egocentric scenes captured in infant homes during everyday activities shows faces to be highly prevalent for infants younger than 3 months and much rarer for infants older than 18 months. In contrast, for younger infants, hands are rarely in view but, for older infants, hands acting on objects (either their own or others') are nearly continuously in view. Infants—through the rewarding dynamic cycles of face-to-face play—generate regularities in behavior and sensory inputs that are prior to and fundamentally

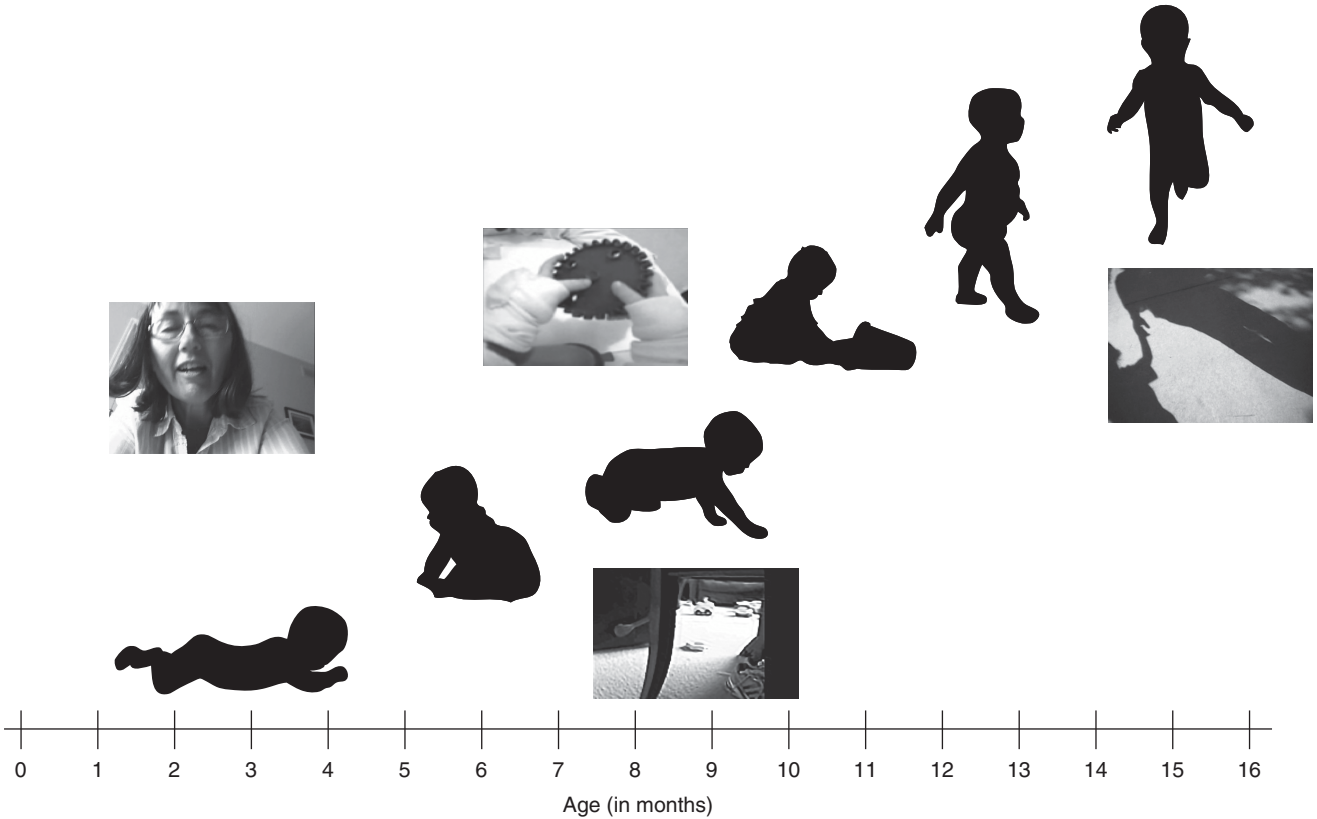


Figure 3.2 Sensory-motor skills and postures change dramatically in the first year and a half of life, with each new sensory-motor achievement leading to new sensory experiences, as illustrated by changing postures of the pictured infants. The images were captured from head cameras worn by a very young infant sitting in an infant seat, a crawling infant, a sitting infant holding a toy, and a walking infant. They illustrate the different views and perspectives provided by changing sensory-motor skills.

different from the regularities generated by toddlers acting and observing the actions of others on objects.

Brain networks change, bodies and what they do change, and the environment and its regularities change in deeply connected ways, with causes and consequences inseparable within the multi-scale dynamics of the brain–behavior–environment network. Theories of how evolution works through developmental processes have noted how evolutionarily important outcomes are often restricted by the density and ordering of different classes of sensory experiences (e.g., Gottlieb, 1991). This idea is often conceptualized in terms of "developmental niches" that provide different environments with different regularities (e.g., West & King, 1987; Gottlieb, 1991) at different points in time. These ordered niches—like a developmental period dense in face inputs or dense in hand inputs—play out in the development of individuals in real time and have their causes and consequences in the dynamic interplay of structural and functional brain networks through the body and in the world across shorter and longer time scales.

Primarily because of limitations in brain imaging technology, there is presently little direct evidence linking these changes in motor development and multisensory input to changes in brain networks in infants and toddlers. However, studies of older children learning to read, write, and compute provide direct evidence of brain networks being modulated by changes in behavior and input statistics (James, 2010; Hu et al., 2011; Li et al., 2013). Literacy acquired during childhood and adulthood is associated with largely similar patterns in structural (De Schotten et al., 2014) and functional (Dehaene et al., 2010) brain networks, underscoring the importance of behavior in creating those changes.

The dynamics of the brain–behavior–environment network—its adaptability, its core properties, and its development—have profound theoretical implications for understanding human cognition. Developmental changes in experiences and in the active sampling of information restructure the input statistics and over time yield changes in brain network topology and dynamics, which in turn support and influence behavior and new experiences. The sources of brain changes relevant to some development can be indirect and overlapping, with handwriting practice influencing reading networks (James, 2010), and reading practice influencing auditory language networks (Monzalvo & Dehaene-Lambertz, 2013). Many behavioral changes—learning to walk, manipulating objects, talking, joint action with others, learning to read—are common and linked with age, and thus seem likely to contribute to the age-related changes being reported in brain network structure and function (Johnson & De Haan, 2015; Pruett et al., 2015). In sum, the changing dynamics of the child's body and behavior modulate the statistics of sensory inputs as well as functional connectivity within the brain, which contribute to developmental changes in the functional and structural networks that constitute the human cognitive system.

Pathways Not Origins

Developmental theorists often refer to the "developmental cascade," and do so most often when talking about atypical developmental processes, such as how

motor deficits and limits on children's ability to self-locomote cascade into the poor development of social skills (Adolph & Robinson, 2015) or how disrupted sleep patterns in toddlers start a pathway to poor self-regulation and conduct disorder (Bates et al., 2002). But the cascade *is* the human developmental process for cognition, typical and atypical alike, and it is the *consequence* of the history-dependence of brain–body–environment networks. Like evolution and culture, new structures and functions emerge through the accumulation of many changes. As William James noted in the earlier quote, we are at each moment the product of all previous developments, and any new change—any new learning—begins with and must build on those previous developments. Because of this fact about development, we submit that the "origins" question (and all talk about nature and nurture) is hopelessly outmoded and must be replaced with the "pathways" question. Rather than ask whether human cognition is innate or learned, we should ask about the nature of the pathway leading to mature human cognition. In biology and embryology, a developmental pathway is defined as the route, or chain of events, through which a new structure or function forms. Thus embryologists delineate the pathways—the details of the chains of events—that lead to the new structures of a neural tube or a healthy liver.

These pathways are evident in cognitive development as well. For example, object handling and manipulation by toddlers generates the sensory input critical to the typical development of three-dimensional shape processing (Pereira et al., 2010), stabilizes the body and head and supports sustained visual attention (Yu & Smith, 2012), makes objects visually large, centered and isolated in toddler visual fields, which support the learning of object names (Yu & Smith, 2012). By simultaneously indicating the direction of the toddler's momentary interests to others, object handling and manipulation invites social interactions and joint attention to an object with social partners (Yu & Smith, 2013). Joint attention to an object with a mature partner extends the duration of toddler attention and may train the self-regulation of attention (Yu & Smith, 2016). But holding and manipulating an object depends on stable sitting (Soska et al., 2010), and stable sitting depends on trunk control and balance (Bertenthal & von Hofsten, 1998). In this way, trunk control is part of the typical developmental pathway for object name learning and for the self-regulation of attention (Smith, 2013).

Just as in embryology, the pathways in behavioral development and the development of brain networks will be complex in three ways that challenge the old-fashioned questions about origins. First, change is multi-causal—that is, each change may be dependent on multiple contributing factors. Second, there may be multiple routes to the same functional end. Third, change occurs across multiple time scales that bring the more distant past into direct juxtaposition with any given current moment. Because developmental pathways may be complex in these ways, the question about necessary or sufficient causes, about "origin," becomes moot. Indeed, in the study of developmental pathways at the molecular level, researchers characterize prior states that set the context for the next event in the chain of developmental events as "permissive to." So, for example, in behavioral development we might say that trunk control is permissive to object manipulation and object manipulation is permissive to object name learning.

What About Cognition?

Traditional views of cognition derive from a separation of mind from the corporeal. In this view, mental life may be partitioned into three mutually exclusive parts: sensations, thought, and action (e.g., Fodor, 1975). Cognition was strictly about the "thought" part and was understood to be amodal, propositional, and compositional, and thus to be fundamentally different from the processes responsible for perceiving and acting, which must deal in time and physics (Pylyshyn, 1980). Although this stance may be weakening given the juggernaut of advancing human neuroscience, many of the theoretical constructs in the study of human cognition and its development have their origins in the traditional view and thus these theoretical constructs are usually understood as strictly cognitive, and neither defined in terms of nor linked to the dynamics of brain, behavior, and world. Thus one might ask: What is the role of these constructs—knowledge, concepts, reference, aboutness, representation, and symbols—within the perspective offered here?

As a first step in his argument for a language of thought, Fodor (1975) made a cogent argument against reductionism. He noted that there could be lawful relations at one level of analysis and lawful relations at another that did not map coherently and systematically to each other, and that, to capture important generalizations, phenomena needed to be studied—and explained—at the proper level of analysis. This is surely correct (although understanding the bridges between levels is also where field-changing and barrier-breaking advances often occur). But can constructs such as *concepts*, *innate*, *learned*, and *core knowledge* be saved by this "level of analysis" move, which states that they are about phenomena at a different level of analysis than the dynamics of brain–behavior–environment networks? Although there are phenomena for which cognitive analyses might be appropriate, we would argue that development is not one of them. As Fodor clearly argued, phenomena need to be understood at the level of analysis that can reveal the explanatory principles for the phenomenon in question. Human cognitive development is a phenomenon of multi-scale, multi-causal change in a very complex system and thus the relevant theory about the development of human cognition must be in these terms.

The old-fashioned dichotomy of origins—learned versus innate—is ill-posed because of the history-dependence and multiple causality of developmental processes—from conception onward. The brain–behavior–environment network begins to form prior to birth when the developing brain begins generating behaviors that affect the fetal environment (e.g., Brumley & Robinson, 2010). The "origins" of every aspect of cognition in a present moment lie in how the individual's developmental pathway has carved out the structural properties and in-the-moment neural activity up to this moment in time.

References

Adolph, K. E., & Robinson, S. R. (2015). Motor development. In R. M. Lerner (Series Ed.) & U. Muller (Vol. Eds.), *Handbook of Child Psychology and Developmental Science*, (Vol. 2: *Cognitive Processes*, pp. 113–57), 7th ed. New York: Wiley.

Adolph, K. E., Tamis-LeMonda, C. S., Ishak, S., Karasik, L. B., & Lobo, S. A. (2008). Locomotor experience and use of social information are posture specific. *Developmental Psychology*, 44(6), 1705.

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, 86(6), 561–5.

Bates, J. E., Viken, R. J., Alexander, D. B., Beyers, J., & Stockton, L. (2002). Sleep and adjustment in preschool children: Sleep diary reports by mothers relate to behavior reports by teachers. *Child Development*, 73(1), 62–75.

Bertenthal, B., & Campos, J. J. (1990). A systems approach to the organizing effects of self-produced locomotion during infancy. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in Infancy Research* (Vol. 6, pp. 1–60). Norwood, NJ: Ablex,

Bertenthal, B., & Von Hofsten, C. (1998). Eye, head and trunk control: The foundation for manual development. *Neuroscience & Biobehavioral Reviews*, 22(4), 515–20.

Bhat, A. N., Landa, R. J., & Galloway, J. C. C. (2011). Current perspectives on motor functioning in infants, children, and adults with autism spectrum disorders. *Physical Therapy*, 91(7), 1116–29.

Brumley, M. R., & Robinson, S. R. (2010). Experience in the perinatal development of action systems. In M. S. Blumberg, J. H. Freeman, & S. R. Robinson (Eds.), *Oxford Handbook of Developmental Behavioral Neuroscience*. Oxford: Oxford University Press, 181–209.

Byrge, L., Sporns, O., & Smith, L. B. (2014). Developmental process emerges from extended brain–body–behavior networks. *Trends in Cognitive Sciences*, 18(8), 395–403.

Chiel, H. J., & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences*, 20, 553–7.

Clark, A. (1998). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Dayan, E., & Cohen, L. G. (2011). Neuroplasticity subserving motor skill learning. *Neuron*, 72(3), 443–54.

Deco, G., Tononi, G., Boly, M., & Kringelbach, M. L. (2015). Rethinking segregation and integration: Contributions of whole-brain modelling. *Nature Reviews Neuroscience*, 16(7), 430–9.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., et al.… & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, 330(6009), 1359–64.

de Schotten, M. T., Cohen, L., Amemiya, E., Braga, L. W., & Dehaene, S. (2014). Learning to read improves the structure of the arcuate fasciculus. *Cerebral Cortex*, 24(4), 989–95.

Destexhe, A. (2011). Intracellular and computational evidence for a dominant role of internal network activity in cortical computations. *Current Opinion in Neurobiology*, 21, 717–25.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, 152, 101–7.

Fodor, J. A. (1975). *The Language of Thought* (Vol. 5). Cambridge, MA: Harvard University Press.

Fontanini, A., & Katz, D. B. (2008). Behavioral states, network states, and sensory response variability. *Journal of Neurophysiology*, 100, 1160–8.

Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9), 700–11.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye-tracking: A new method to describe infant looking. *Child Development*, 82(6), 1738–50.

Gottlieb, G. (1991). Experiential canalization of behavioral development: Results. *Developmental Psychology*, 27(1), 35.

Hagmann, P., Grant, P. E., & Fair, D. A. (2012). MR connectomics: A conceptual framework for studying the developing brain. *Frontiers in Systems Neuroscience*, 6, 43.

Han, F. et al. (2008). Reverberation of recent visual experience in spontaneous cortical waves. *Neuron*, 60, 321–7.

Harmelech, T., Preminger, S., Wertman, E., & Malach, R. (2013). The day-after effect: Long term, Hebbian-like restructuring of resting-state fMRI patterns induced by a single epoch of cortical activation. *Journal of Neuroscience*, 33(22), 9488–97.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–54.

Hu, Y., Geng, F., Tao, L., Hu, N., Du, F., Fu, K., & Chen, F. (2011). Enhanced white matter tracts integrity in children with abacus training. *Human Brain Mapping*, 32(1), 10–21.

James, K. H. (2010). Sensori-motor experience leads to changes in visual processing in the developing brain. *Dev. Sci.*, 13, 279–88.

James, K. H. & Engelhardt, L. (2012). The effects of handwriting experience on functional brain development in pre-literate children. *Trends in Neuroscience and Education*, 1, 32–42.

James, K. H., Jones, S. S., Smith, L. B., & Swain, S. N. (2014). Young children's self-generated object views and object recognition. *Journal of Cognition and Development*, 15(3), 393–401.

James, W. (1950). *The Principles of Psychology* (Vol. I). New York, NY: Dover. (Original work published 1890)

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PloS One*, 10(5), e0123780.

Johnson, M. H., & De Haan, M. (2015). *Developmental Cognitive Neuroscience: An Introduction*. Hoboken, NJ: John Wiley & Sons.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, 85(4), 1503–18.

Lewis, C. M., Baldassarre, A., Committeri, G., Romani, G. L., & Corbetta, M. (2009). Learning sculpts the spontaneous activity of the resting human brain. *Proceedings of the National Academy of Sciences*, 106(41), 17558–63.

Li, Y., Hu, Y., Zhao, M., Wang, Y., Huang, J., & Chen, F. (2013). The neural pathway underlying a numerical working memory task in abacus-trained children and associated functional connectivity in the resting brain. *Brain Research*, 1539, 24–33.

Luo, C., Guo, Z. W., Lai, Y. X., Liao, W., Liu, Q., Kendrick, K. M., et al. & Li, H. (2012). Musical training induces functional plasticity in perceptual and motor networks: Insights from resting-state FMRI. *PLoS One*, 7(5), e36568.

Mackey, A. P., Singley, A. T. M., & Bunge, S. A. (2013). Intensive reasoning training alters patterns of brain connectivity at rest. *Journal of Neuroscience*, 33(11), 4796–803.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.) (2013). *Machine Learning: An Artificial Intelligence Approach*. Berlin: Springer Science & Business Media.

Monzalvo, K., & Dehaene-Lambertz, G. (2013). How reading acquisition changes children's spoken language network. *Brain and Language.*, 127, 356–65.

Northoff, G., Qin, P., & Nakao, T. (2010). Rest-stimulus interaction in the brain: A review. *Trends in Neurosciences*, 33(6), 277–84.

Pereira, A. F., James, K. H., Jones, S. S., & Smith, L. B. (2010). Early biases and developmental changes in self-generated object views. *Journal of Vision*, 10(11), 22–32.

Piaget, J. (1952). *The Origins of Intelligence in Children* (Vol. 8, No. 5, pp. 18–1952). New York, NY: International Universities Press.

Pruett, J. R., Kandala, S., Hoertel, S., Snyder, A. Z., Elison, J. T., Nishino, T., et al. & Adeyemo, B. (2015). Accurate age classification of 6 and 12 month-old infants based on

resting-state functional connectivity magnetic resonance imaging data. *Developmental Cognitive Neuroscience*, 12, 123–33.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(01), 111–32.

Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, 14, 180–90.

Sadaghiani, S., Poline, J. B., Kleinschmidt, A., & D'Esposito, M. (2015). Ongoing dynamics in large-scale functional connectivity predict perception. *Proceedings of the National Academy of Sciences*, 112(27), 8463–8.

Sampaio-Baptista, C., Khrapitchev, A. A., Foxley, S., Schlagheck, T., Scholz, J., Jbabdi, S., et al. & Kleim, J. (2013). Motor skill learning induces changes in white matter microstructure and myelination. *Journal of Neuroscience*, 33(50), 19499–503.

Smith, L. B. (2013). It's all connected: Pathways in visual object recognition and early noun learning. *American Psychologist*, 68(8), 618.

Smith, L. B., & Sheya, A. (2010). Is cognition enough to explain cognitive development? *Topics in Cognitive Science*, 2(4), 725–35.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*, 14(1), 9–17.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, 16(3), 407–19.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: Motor skill acquisition facilitates three-dimensional object completion. *Developmental Psychology*, 46(1), 129.

Sporns, O. (2011). *Networks of the Brain*. Cambridge, MA: MIT Press.

Stiles, J. (2008). *The Fundamentals of Brain Development: Integrating Nature and Nurture*. Cambridge, MA: Harvard University Press.

Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron*, 65(2), 280–90.

Taubert, M., Lohmann, G., Margulies, D. S., Villringer, A., & Ragert, P. (2011). Long-term effects of motor training on resting-state networks and underlying brain structure. *NeuroImage*, 57(4), 1492–8.

Thelen, E., & Smith, L. B. (1994) *Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT press.

West, M. J., & King, A. P. (1987). Settling nature and nurture into an ontogenetic niche. *Developmental Psychobiology*, 20(5), 549–62.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy*, 13(3), 229–48.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, 125(2), 244–62.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye–hand coordination. *PloS One*, 8(11), e79659.

Yu, C., & Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. *Current Biology*, early online, doi: http://dx.doi.org/10.1016/j.cub.2016.03.026

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, 16(6), 959–66.

Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: Neuroimaging changes in brain structure during learning. *Nature Neuroscience*, 15(4), 528–36.

4 The Metaphysics of Developing Cognitive Systems

Why the Brain Cannot Replace the Mind

Mark Fedyk and Fei Xu

1. Introduction

Where is the mind? One of the aims of cognitive science is to answer this question and in so doing provide an answer more substantive than the assertion that parts of it are somewhere behind the eyes and between the ears and that it involves, somehow, the brain. The chapter offers an answer to this question: We shall argue that the mind is a computational network that occupies a functional location in a more complex causal system formed out of various distinct but interacting neurological, physiological, and physical networks. The mind therefore has a functional location.

The idea of a functional location is important because only static entities have relatively fixed spatio-temporal properties. By their very nature, dynamic networks have constantly changing locations, even though states of these networks can usually be spatially located. Nevertheless, a network can be located by specifying how its endogenous processes and states interact with exogenous processes and states—as well as exogenous networks, systems, or, indeed, any other kind of causal process. The network itself can be located by describing, at least in part, its *functional role* within a larger web of cause and effect; again, this is to give the functional location of the network.

We begin with the distinction between functional and spatial location because Smith, Byrge, and Sporns (this volume) offer a different answer to the question of where the mind is located. According to them, the mind—in the non-reductive, Fodorean sense—is not real because it cannot be functionally located in relation to the brain. Their argument for this is novel and ingenious; they reason as follows: If cognition is real and cognition is computation, then the cognitive system must be composed out of a set of static states. But all of the parts of the brain are dynamic processes and entities. So, for any states of the cognitive system to also be components of the brain, they too would have to be dynamic; yet, since cognitive states are computational states, they must be static. It follows that there are no cognitive states. The brain is constituted in a way that prevents the mind from occupying a functional location amongst its many different neural processes and connective networks.

We disagree with Smith and her co-authors. There is a coherent sense in which systems composed of dynamic processes and computational systems can be

co-instantiated (Siegelmann & Sontag, 1995; Whittle 2007). We therefore reject the major premise of Smith et al.'s argument. Indeed, there are many examples of this kind of co-instantiation. For instance, every electronic device with a computer chip in it is an example of a computational system (the electrical network running on the chip) that is co-instantiated with a physical system (the circuits etched onto different physical components linked together by a complex of wires and channels). More esoteric examples are provided by phenomena like membrane computation (roughly, those things which implement P-systems; cf. Păun & Rozenberg, 2002), biological processes that implement cellular automata, for example colour patterns of certain cephalopod molluscs (Packard, 2001; but see also Koutroufinis, 2017), and models of gene expression and regulation that conceptualize both as a form of "natural" computation (Istrail et al., 2007). These examples make it clear that it is neither logically impossible nor scientifically improbable that the mind is a computational network co-instantiated with a number of dynamic neural networks.

And yet, because almost all of the metaphysical theory of the development of dynamic systems that Smith and her co-authors offer is one that friends of cognition can, and should, accept, we shall use this chapter to pursue a deeper response to Smith et al. Our primary aim will be to extend the theory of the metaphysics of cognitive development that Smith et al. provide so that the extended theory can be used to give a (partial) account of the mind's functional location. Our extension of Smith et al.'s dynamic systems theory will also allow us to clarify exactly why there is no inference from the dynamism of the brain's networks to the non-existence of a cognitive system. But, in establishing this clarification, we will also prepare the ground for our secondary aim, which is to establish support for an argument that shows why, given a choice between our extended view of cognitive development and Smith et al.'s comparatively restricted view, the extended theory should be favored. Here, the argument is simple: Our extended view can explain more scientific data than Smith et al.'s restricted theory—for that reason, the extended view should be preferred.

In more detail, here is how we shall proceed. We will start from an abstract examination of the metaphysics of dynamic systems (section 2.0). We begin here because many of the most scientifically interesting dynamic systems have two salient ontological features. First, they are made up of component systems and mechanisms that are organized at different levels of energy and constructed out of different forms of energy. Second, many of these systems can realize (sometimes extraordinarily complex) functions. Thus, our initial goal is to provide an explanation of these two observations. To do this, we introduce the concepts *causal buffering* and *metaphysical transduction* to explain how dynamic systems can be made up of systems that are able to pass information between themselves, thereby either establishing or maintaining the function of the system, without also transmitting so much energy as to cause the overall system to break apart. Then, we turn to innateness. Smith et al. are skeptical that the concept of innateness is compatible with a dynamic systems worldview, but we show (section 3.0) that there is a way of defining innateness in terms of developmental essentiality that is not only compatible with this worldview but also helpful for explaining how new systems can emerge as components, or byproducts, of existing dynamic systems. To wit: The

existence of certain innate causal buffers (section 4.0) and certain innate metaphysical transducers (section 5.0) provides an elegant explanation of how mind can be a computational system that is co-instantiated with the brain's dynamic neurological networks. This conclusion provides us with the premise we need to argue that consilience considerations favor our extended view as compared to Smith et al.'s more restricted view (section 6.0). Lastly, we return to the question of the functional location of the mind in this chapter's final brief conclusion (section 7.0).

2. The Metaphysics of Systems of Systems

Smith et al. review in impressive detail the many ways that interactions with the proximate environment can both dynamically alter, and drive increases in the structural complexity of, various neural and physiological networks. Brain, body, and environment (BBE, hereafter) are constantly causing changes to one another. We think that the cognitive mind needs to be added to this list of interacting networks. But in this section we will focus only on the metaphysical architecture of the complex, dynamic network formed out of brain, body, and environment—since it is not possible to disagree with Smith et al.'s contention that the BBE network plays a central role in structuring virtually all levels of development, including cognitive development.

Our starting point is an observation about the physical integrity of BBE networks—namely, that the integrity or stability of a BBE network cannot be taken for granted. There are no laws of nature which create or necessitate these networks, after all; BBE networks are not the automatic byproducts of nomological necessities. Instead, the physical integrity of a BBE network is normally a causal byproduct of the BBE network's endogenous structure. Yet a durable, functioning BBE network is nevertheless something like an unplanned, accidental circus act: Things as massive as an elephant and as small as a mouse, as ephemeral as a soundtrack and as abstract as a set of linguistic descriptions and commands, all must find a way of interacting in a sustained, coordinated, and causally integrated fashion.

The analogy is imperfect, of course, partly because it isn't clear what (beyond entertainment) the functions of circus acts include—but the analogy is nevertheless helpful. The analogy is helpful because it calls our attention to the fact that the components of a BBE network must interact only in very specific ways for the BBE network as a whole to both maintain its physical integrity and to consequently realize complex functions such as learning, progressively increasing task proficiency, or even just contextually appropriate behavior. This in particular is puzzling. BBE networks, like circus acts, can be made out of component subsystems that are themselves organized at very different levels of energy and constructed out of very different physical formats. But, unlike circus acts, BBE networks seem to be capable of degrees of a functional self-regulation, which requires the component systems of a BBE network to be able to share information (or at least signals) with one another. Consequently, the fact that a BBE network can maintain its endogenous structure—and thereby preserve its physical integrity as well as its

functional capabilities—gives rise to two overlapping metaphysical questions. First, how is it that component physical systems of BBE networks that are frequently organized at very different magnitudes of energy are able to be components of the same overall system? Second, how is it that the component systems of any BBE network are able to send information between themselves without also transmitting so much energy as to cause the BBE network itself to break apart? Or, putting the questions a bit more abstractly: What is it about the metaphysics of BBE networks that explains both why they can maintain their physical integrity and why they can realize various complex functions?

We think the work on the metaphysics of mechanisms which has occurred in the philosophy of science over the last two decades contains the seeds of an answer to these two metaphysical questions (Cummins, 2000; Tabery, 2004; Craver & Bechtel, 2007; Glennan, 2017; Matthews & Tabery, 2017; Love, 2018). This is because, considered very abstractly, a BBE network is a system formed out of a number of complex component mechanisms, which themselves frequently take the form of causal systems. Indeed, a useful way of simplifying the import of this literature for our account of the metaphysics of BBE networks is to see this literature as providing the impetus for drawing a very general distinction between mechanisms and systems of mechanisms (cf., Craver, 2001)—or, as we shall say, a very general distinction between *simple systems* and *systems of systems*, as this choice of words makes our terminology a bit more straightforward.

Here is how we want to define this distinction. First of all, simple systems are closed networks of causally interacting components, where the energy transferred between the components is of roughly the same magnitude. This fact can explain why a simple system maintains its physical integrity: It is hard, or impossible, for the simple systems' endogenous causal processes to acquire sufficient power to be able to break or destroy the system as a whole. Most things in nature, though, are not simple systems. They are usually (dynamic, complex, and/or emergent) systems of systems. A system of systems is a system whose components are systems that would be simple systems if it were possible to extract them from the system of systems without breaking the simple system itself. The crucial difference, then, is that, unlike a simple system, a system of systems can have component systems that are organized (or are stable) at substantially different magnitudes of energy. Thus, the universe is one extremely big system of systems—and so too are all BBE networks. Furthermore, our view that the mind is a computational system is, metaphysically speaking, the proposal that the mind be conceptualized as one simple system amongst others in the system of systems formed by any real-life BBE network.

The question before us now, however, is only: What is it about the metaphysics of BBE networks, *given that they are systems of systems and not simple systems*, that explains how these systems maintain both their causal integrity and their ability to realize any number of different functions? An example will help us answer this question. Suppose that there exists a BBE network made of a child bouncing a red ball in a well-lit room with no one else present. This system of systems has amongst its constituent systems causal processes manifest in the forms of kinetic energy (the bouncing ball), radiant energy (photons reflected from the ball to the

retina), and chemical energy (hydrolysis in the retina). Each of the component systems is stable and able to contribute to the smooth functioning of the system of systems despite the fact that the energy involved in each of these component systems is multiple orders of magnitude greater or smaller than in the other systems. Put somewhat baldly: If, for instance, the total mechanical energy in the bouncing ball were transmitted directly into the brain, it would destroy at least the cortex. The integrity of a system of systems, then, depends on the ability to shield, insulate, protect, or otherwise buffer the component simple systems from one another. Or, put more generally, systems of systems are stable because of *causal buffering*. In this case, causal buffering is provided by, inter alia, the flexibility of the child's arm, the child's perceptual coordination capacities, and perhaps even the child's skull itself.

Yet, it is important to observe that perfect causal buffering—causal buffers that block *all* energy transfer between systems—would prevent the system of systems from realizing even the simplest of functions. If no energy whatsoever could follow a loop running between the child's brain and the ball, then bouncing the ball would be impossible. *A fortiori*: A body, itself a system of systems, would not be able to maintain homeostasis if it were impossible for its component simple systems to interact with one another. So, in addition to causal buffering, systems of systems must allow component systems to share information—which we will call *metaphysical transduction*. In our example, metaphysical transduction is realized by the mechanisms which implement, inter alia, the child's proprioception of her arm's location, the mechanisms which convert electromagnetic radiation into biochemical energy via the process of visual phototransduction and eventually generate the child's input visual cues, feedback from different clusters of striated muscles, and information from afferent neurons in the child's hands—all of which ensure that the child maintains a sense of the ball's location relative to her own body and its own path of spatial movement.

It is easy to find examples of causal buffers and metaphysical transducers that respectively hold together systems of systems and permit the system as a whole to have any number of functions. Take, for example, any commercially produced car. When running, the engine produces vibrations which, if not absorbed, would shake the engine apart. The engine is buffered against itself by, inter alia, ensuring that its heaviest moving parts are in mechanical balance, increasing the mass of the engine block, placing the camshaft above the combustion chamber, and mounting the engine to the chassis using extremely durable rubber vibration dampeners. But vibration isn't the only kind of energy that threatens the physical integrity of the car. A separate array of buffers is used to control the heat generated by the engine; in most cases, this is the function of the radiator, but the radiator cannot perform its function if the oil which provides lubrication is not also buffering parts of the engine from the damaging effects of heat that is caused by friction.

Cars, of course, are meant to be driven; the steering system provides us with an elegant example of an interlocking chain of metaphysical transducers connecting the steering wheel to the front wheels. In a rack-and-pinion layout, for instance,

the steering wheel turns a column which then turns a pinion gear that is meshed with a linear gear fixed atop a rack—the net effect of which is to turn the radial motion of the steering wheel into the horizontal motion of the rack itself. The rack is connected by way of tie rods (which act as both causal buffers and metaphysical transducers) to the king pin, which turns the wheels. If the strength of the driver is insufficient to produce enough torque to turn the wheels, the steering system will have hydraulic or electric actuators which amplify the steering inputs produced by the driver. Similar chains of metaphysical transducers connect the gas pedal with the throttle, the brake pedals with the brakes, and the numerous electronic control units (the PCU, TCU, and so on) with various subsystems endogenous to the system of systems that is any car. (And to foreshadow: We think it is not accidental that the computational systems embedded in the car's ECU, for instance, dramatically increase the number of scenarios that the car's engine can operate *optimally* within.)

So, just as the causal buffers and metaphysical transducers built into a car explain why the car does not explode, melt, or shake itself to bits, and also how signals are able to pass between different component systems such that the car is able to drive, so too will there be parts of BBE networks that function as causal buffers and metaphysical transducers, and which therefore explain why BBE networks can maintain their causal integrity while realizing different functions as complex as learning, or even comparatively simpler functions, such as planning, playing, mindreading, or absentmindedly bouncing a ball.

3. Innateness as Developmental Essentiality

We have begun with a very general analysis of the metaphysics of systems of systems because it leads us to an important insight into the architecture of the cognitive system: Since our hypothesis is that the cognitive system is a simple system within a larger system of systems, it too should have its own causal buffers and metaphysical transducers. Moreover, at least some of the relevant buffers and transducers must be innate—for it is the innateness of at least some of the mind's buffers and transducers that explains how the cognitive system can develop. The cognitive system, just like the other component systems of BBE networks, does not just appear out of nowhere. All of these systems are causal byproducts of "precursor" systems, and our conception of innateness provides an explanation of how this is possible.

However, this line of reasoning depends upon a new conception of innateness; the difference in meaning between how we shall use the concept and how it is customarily used in philosophy, biology, and psychology is large enough to warrant a formal definition. Accordingly, we will start this section with an explanation of our conception of innateness, one that is designed to fit within the dynamic systems worldview, before turning to an explanation of how this concept can be used to explain the development of new simple systems within a larger system of systems.

The concept of innateness is customarily used to denote traits that are in some sense fixed or immalleable, such that what makes a trait innate is something like its invariance under different kinds of developmental, genetic, or environmental

pressures.[1] We want to use, instead, a concept of innateness that expresses the idea that traits are innate because they are developmentally essential, and, for that, not *necessarily* fixed and immalleable over the full temporal duration of the system in which the trait is a part. This idea can be unpacked by returning to the question of how a new simple system can develop within an existing system of systems. Indeed, the biological world provides countless examples of systems of systems that have amongst their functions the power to, from time to time, produce a new simple system. When this happens, there will be a period of time during which the "parent" or "precursor" system overlaps with, and therefore shares some of the parts of, the "child" or "derivative" system—life, after all, does not begin or end; it is only selectively transmitted.

The idea, then, is that innate traits will be the components of the "child" systems that are byproducts of the operation of a "parent" system, which can, after a certain amount of time, become elements of the "child" system. These traits will also be essential to stabilizing the functions of the "child" system *because* they provide, at least initially, the causal buffering and metaphysical transduction needed for the "child" system to separate from the "parent" system, all without disrupting the functions realized by the overarching system of systems. Or, to put the same idea a different way, what makes a trait innate is time and system-relative: The innate traits of a simple system are just those traits which are amongst the initial parts of a new simple system and which are causally necessary for the new system to become a discrete system when the system is itself the causal byproduct of the operation of other simple systems in a system of systems.

That is the abstract outline of the concept; we can further clarify it by defining it explicitly. Thus, according to this new concept, a trait or mechanism is innate if and only if:

- The trait is *developmentally essential* to at least one of the simple systems that it is a part of; without this trait, the system in question cannot come into existence.
- The trait exists, proximately speaking, because it is a causal byproduct of one of the systems that it is *not* a developmentally essential part of.
- For at least a meaningful period of time after the trait comes into being, the trait can only modify, but not be modified by, causal processes that are endogenous to at least one of the systems that the trait is a developmentally essential part of.

Now, since our intent is only to use *this* concept of innateness—not to argue that it is something like the one *single* true concept of innateness—it will suffice to justify our characterization of certain cognitive mechanisms as innate using this concept by finding evidence that our tripartite definition is not empirically vacuous. Consider, thus, the genome of any organism—it will be innate by our definition. In sexually reproducing species, the processes of meiosis and fertilization that create a unique set of chromosomes occur in physiological systems that almost always lose the set of chromosomes as a part, but the same set of chromosomes is a developmentally essential component of a great number of different physiological

systems. Finally, while complex feedback loops regulate the causal powers of a genome throughout much of its existence (Jablonka et al. 2014), the earliest stages of cell growth and differentiation are mostly biochemical effects of the genome itself (cf., Reik et al. 2001; Mizushima & Levine 2010).

As noted above, this concept of innateness allows us to say that some trait is innate *relative* to a particular system, but not innate relative to another system—even if the trait, for some non-trivial period of time, is a part of the second system. This is important because it is not possible for new systems to develop within existing systems of systems without the new system sharing most, if not all, of its component mechanisms with a precursor system for at least a short period of time. So, this conception of innateness allows us to distinguish between a trait being a byproduct of a parent system and thus not innate relative to the parent system yet nevertheless being developmentally essential to a child system and thus innate relative to this second system.

Because of its ability to mark out this distinction, this concept of innateness allows us to express the idea that the ability of new simple systems to develop within systems of systems is only possible because certain causal buffers and metaphysical transducers are innate—even if some of the buffers and transducers are either effects, or even parts of, the precursor system. Put more concretely, the idea is that, just as some of the brain's innate traits (e.g., the blood–brain barrier) explain how it emerges as a stable simple system within a system of systems constituted by bodily and environmental networks, some of the mind's innate traits can *explain* how a computational system emerges as a distinct system within a system of systems.

What might these innate transducers and buffers be? Amongst the transducers must be mechanisms that are able to convert streams of different non-cognitive signals into cognitive information, and also mechanisms which convert cognitive information into non-cognitive signals. Amongst the causal buffers, there must be mechanisms that permit a computational system to remain sufficiently insulated from potentially interfering forces for it to remain co-instantiated with the brain's neurological networks and systems.

Ultimately, we are most interested in the former—since digging a bit deeper into empirical theories of the mind's innate metaphysical transducers might lead to the intriguing conclusion that there are probably a number of innate concepts. However, before turning directly to the question of whether there are any innate concepts, we want to first return to Smith et al.'s argument that the dynamism of BBE networks is a reason to be skeptical of the existence of cognition. Exploring what it means to say that the cognitive system is co-instantiated with a variety of other systems sheds some light on what some of the mind's innate causal buffers may be.

4. Co-instantiation, Computation, and Degeneracy

Smith et al. are skeptical that the discrete and static states of any computational system can be functionally located within the brain. Turing also dealt with this problem. In the paper largely responsible for introducing the computational theory of cognition, Turing offers the following observations:

> [Discrete state machines] are the machines which move by sudden jumps or clicks from one quite definite [i.e. static] state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously. But there are many kinds of machines which can profitably be *thought of* as being discrete state machines. For instance in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them.
>
> (Turing, 1950, p. 439, emphasis in the original)

Turing's uses of "thought of" and "convenient fiction" are usefully ambiguous. One interpretation of what Turing means to say is that discrete state machines, and therefore digital computers, do not exist *simpliciter*. But this interpretation is contradicted by Turing's subsequent assertion that it is possible to build discrete state machines that are digital computers, but only if the physical system out of which both are built reduces to almost nil the chance that the continuously moving, dynamically interacting physical parts will cause the computer to depart from its programming. This suggests that the alternative interpretation which more accurately captures Turing's intended meaning is one which reads him as saying that discrete state machines, and therefore digital computers, and dynamic physical systems can be (and frequently are—think of all of the switches you have used today) co-instantiated. And one way of unpacking the meaning of the co-instantiation thesis is seeing that it implies that discrete state machines cannot actually be built *simpliciter*. We cannot build a physical system that is also a digital computer and which has exactly zero chance of departing from its programming. Nevertheless, we can build physical devices the operations of which are so *extremely* well-aligned with the operations of hypothetical zero-error computers that what gets built is a physical system that is co-instantiated with a non-zero-error (and therefore quasi-) computer.

Thus, there are two simple systems that are co-instantiated in any real-world digital computer: the continuous (or dynamic) physical components of the system and the static computational components of the (non-zero-error) computer system itself. The key point, then, is that the former so closely mirrors the operations of an entirely hypothetical zero-error computing machine that nothing is lost by thinking of the real non-zero-error quasi-computer system as if it is really the hypothetical zero-error computing machine. Turing is denying only the physical reality of zero-error digital computers, and asserting that non-zero-error computers can be co-instantiated with all sorts of physical systems.

This shows that it is conceptually possible for computational systems to be co-instantiated with larger dynamic systems. But this does not completely answer the question of how the static states of a computational network can be co-instantiated with the dynamic networks of the brain. The crux of the issue is that the dynamism of neural networks makes brains highly variable, both across individuals and over meaningful periods of developmental time. As Smith et al. stress, the dynamic properties of different brain networks, and the massively differential

impact that variations in both behavior and environment can have on brain development, mean that patterns of neural connectivity are extremely variable from individual to individual, from behavioral context to behavioral context, and from environmental context to environmental context. Turing's observation that static systems can be co-instantiated with dynamic systems does not help us address the question of how a static system with the same functionality (say, implementing the inferential processes that infer edges from stereopsis) can be co-instantiated with a very large set of inherently different connective systems.

Put more precisely, however, this problem really just is the problem of explaining how there can be coincident causal realization of two systems without there being a homomorphism between the structures of the two systems. And this problem is solved by evidence that a one-to-many relationship holds between the functional organization of the computational mind and different neural networks. This, in turn, amounts to evidence that the brain has substantial amounts of what Edelman (1987; Tononi et al., 1999; Edelman & Gally, 2001) calls *degeneracy*:

> Degeneracy is the ability of elements that are structurally different to perform the same function or yield the same output. Unlike redundancy, which occurs when the same function is performed by identical elements, degeneracy, which involves structurally different elements, may yield the same or different functions depending on the context in which it is expressed. It is a prominent property of gene networks, neural networks, and evolution itself. Indeed, there is mounting evidence that degeneracy is a ubiquitous property of biological systems at all levels of organization.
>
> (Edelman & Gally, 2001)

Importantly, degeneracy is an empirical concept. With his collaborators, Edelman has shown that there are high levels of degeneracy in the brain's neural networks—a result that has been used by several subsequent researchers to explain how different computational functions can be co-instantiated with the different forms of connectivity inherent to any living brain (Eliasmith & Anderson, 2004; Eliasmith, 2007; Park & Friston, 2013).[2] Indeed, in an earlier article, Smith herself recognizes the importance of degeneracy: "The notion of degeneracy in neural structure means that any single function can be carried out by more than one configuration of neural signals and that different neural clusters also participate in a number of different functions" (Smith, 2005, p. 290). And finally, the brain is *innately* degenerate: Degeneracy is developmentally essential for the emergence of all neurological networks which share at least some physiological resources.

It should therefore be unsurprising that digital computers provide another example of degeneracy, albeit at the level of instruction set architecture. A microprocessor's instruction set specifies what computational functions that processor can perform—familiar architectures include the original x86 specification, extensions to it like SSE and AMD64, and the growing family of the ARM specifications. The circuits etched into silicon which implement these instruction sets can be radically different: There are thousands of very different microprocessor chips which have, for instance, the function of implementing the 32-bit variant of

x86. (Technically, these circuits are non-zero-error implementations of the relevant instruction sets.) Transistors are a different example of a simpler form of degeneracy in electrical engineering: There are now thousands of different physical systems out of which transistors can be built. Finally, most field programmable gate arrays provide us with examples of degenerate computational systems that are co-instantiated with highly dynamic physical systems.

What's more, these observations show us something interesting about the notion that there are *literally* levels of analysis or levels of explanation that both sit within the domain of psychology and also separate psychology from other fields in the cognitive and behavioral sciences. The idea, put roughly, is that the theories of one field will not reduce to the theories of another field because they are about different metaphysically discrete *layers* or *planes* of reality. The disciplinary structure of the cognitive and behavioral sciences mirrors the layered organization of nature, or at least the cognitive and behavioral parts of nature: Each discipline studies a horizontally organized plane formed of phenomena that interact with phenomena on its plane only, and interact according to laws or generalizations that apply to that plane only (cf. Fodor, 1974; Fodor, 1997). Planes that are below provide some kind of ontological or metaphysical support for planes that are above, but, despite this, the laws of a lower or higher plane do not apply to any phenomena except those which occur on the plane itself. And there aren't "bridge laws" either—these would be "vertical" laws that connect the projectible terminology of one disciplinary vocabulary with the vocabulary of another discipline, where the vocabularies apply to different planes, and where the bridge laws serve to establish synonymous definitions for some of the concepts from the first discipline in terms of concepts from the second discipline. This is, we suggest, a rough sketch of the popular picture in the philosophy of mind (cf., Bermudez, 2007).

Yet, it is not a picture that we can wholly endorse. We are happy with the notion that there are levels of analysis, so long as this is taken only as a methodological metaphor (Boyd, 1993), i.e., a metaphor calling attention to certain facts from the research history of the cognitive sciences—facts such as that you cannot do all of the work that is interesting and projectible in cognitive psychology using the methods and concepts of neuroscience (and vice versa). But we have to stop at the point at which the metaphor of metaphysical levels of analysis gets turned into a theory of the fundamental ontological organization of reality according to which reality is literally organized into planes that have some kind of inherent or objective top-to-bottom geometry which allows us to order these planes in relation to one another. We cannot accept this theory—again, despite its apparent popularity amongst some philosophers of mind (cf., Kim, 1990; McLaughlin & Bennett, 2018)—because our commitment to the co-instantiation of computational systems with physical systems means that we are committed to a host of complex causal interactions between any (non-zero-error) computational system and the physical system with which it is co-instantiated. These causal interactions must occur in order for the computational system to be appropriately causally buffered, and for the computational system to play a role, with the help of certain metaphysical transducers, in supporting the functions realized by whatever overarching system of systems the computational system is a constituent of. Or,

to put the same idea another way, we think that nature is a single plane—that of all physical stuff—and that, in some sense, almost everything is co-instantiated with something else. But there are also naturally occurring systems—and systems of systems, and systems of systems of those systems, etc.—that are sustained by all sorts of different kinds of causal buffering, and it is these complexes of causal buffers which, in turn, explain the persistence of systems like the cognitive system, but also the body's various physiological systems, and even large-scale systems like a national economy or a whole ecosystem. We think that scientific disciplines frequently succeed in their efforts to construct conceptual schemes which are mostly proprietary tools for referring to the endogenous causal activity of these systems—and that this is enough to explain why cognitive psychology is (literally) about a different set of phenomena than, inter alia, cognitive neuroscience, neuroanatomy, neurophysiology, and so on. Accordingly, we do not think that the recognition that the cognitive sciences are autonomous relative to one another implies a metaphysical theory according to which there are layers of reality organized in some top-to-bottom fashion according to some *a priori* metric of fundamentality (cf., Davidson, 1973).

But let us get back to the specific case concerning the co-instantiation of a (non-zero-error) computational system with different physical systems. This specific co-instantiation shows that it is *scientifically plausible* to adopt the position that there need not be a homomorphism between the structures of two or more physical systems that, in turn, are co-instantiated with computational systems that have the same function. Evidence that the brain's neural networks are extremely dynamic supports no meaningful *a priori* conclusions about the possible structures of any systems, computational or otherwise, that are co-instantiated with these networks. For all we know, the clusters of properties which constitute an interesting kind at one level of causal interaction (cognitive computation) may, at another level of causal interaction (neural connectivity), form no interesting clusters at all. As a purely conceptual matter, then, it is possible for a computational system to be co-instantiated with a dynamic neurological system. This conclusion is enough to refute the major premise of Smith et al.'s argument.

More important, however, is the observation that co-instantiation and degeneracy also explain some aspects of how the cognitive system is causally buffered, which thereby explains why cognitive psychology is an autonomous discipline. Co-instantiation means that the physical mechanisms and processes which (sometimes literally) insulate different neurological networks from external disruption and interference can confer the same benefit upon the cognitive system, too: Given that they are co-instantiated, whatever mechanisms buffer the brain's non-cognitive neurological networks also buffer the brain's cognitive networks. Furthermore, the brain's inherent degeneracy explains why dynamic changes in neurological networks need not induce changes in computational function: Degeneracy explains how there can be an island of (relative) computational stability in a sea of (again, relative) constant neurological and physiological change. Thus, degeneracy buffers computational function from change caused by ongoing patterns of change in the physical systems which realize the relevant

(non-zero-error) computational systems. Or, put another way, the brain's *innate* degeneracy is likely amongst the most important causal buffers for the computational system that is any human mind.

5. The Innateness of the Initial Conceptual Repertoire

We turn now to the issue of whether there are innate concepts. We know that the mind must contain its own innate metaphysical transducers—the function of which is of course to convert into cognitive information various non-cognitive signals available to the mind as the output of the body's own suite of sensory transducers. And, here, it is important to keep metaphysical questions and scientific questions separate, because the conclusion that the mind must contain a number of innate metaphysical (or cognitive) transducers does not provide an answer to the many and much more difficult scientific questions about the specific empirical form that these transducers take.

That said, there are many sophisticated theories of the empirical form of the mind's innate cognitive transducers to choose from (cf., Samuels, 2000; Marcus, 2006; Smolensky & Legendre, 2006; Heyes, 2018; Schulz, 2018). We believe that one of the most conservative scientific accounts of the mechanisms likely responsible for the earliest forms of non-cognitive-to-cognitive-transduction comes from Susan Carey. According to Carey, the relevant mechanisms should be thought of as dedicated input analyzers: "A dedicated input analyzer computes representations of one kind of entity in the world and only that kind. All perceptual input analyzers are dedicated in this sense: the mechanism that computes depth from stereopsis does not compute color, pitch, or number" (Carey 2011a, p. 451). If (as we have argued is the case) some of these input analyzers are innate, it follows that there must also be a handful of innate concepts as well, namely whichever concepts are embedded in these dedicated input analyzers and which allow the analyzers to produce as output information that is richer—because it contains more structure, or is more abstract, or refers to an unobserved kind or process—than the information that is the input to the analyzer. Whatever else they are, concepts are what represent unobserved or unobservable properties and kinds.

There is compelling evidence that young children have abstract concepts for objects, agents, numbers, and probably also causes (Xu & Carey, 1996; Wang & Baillargeon, 2006; Carey, 2011a; Baillargeon et al., 2012). Our proposal, then, is that the hypothesis that there are innate input analyzers dedicated to generating conceptual representations of objects, agents, numbers, and causes represents the most empirically plausible way of cashing out the more abstract metaphysical conclusion that some metaphysical transduction must take place in order to transform non-cognitive information into cognitive (i.e., computationally tractable) information.

Carey characterizes the mind's innate conceptual resources the following way: "What I mean for a representation to be innate is for the input analyzers that identify the represented entities to be the product of evolution, not the product of learning, and for at least some of its computational role to also be the product of

evolution" (Carey, 2011a, p. 453). But she also resists defining innateness in terms of static, fixed, or non-malleable properties:

> Some innate representational systems serve only to get development started. The innate learning processes (there are two) that support chicks' recognizing their mother, for example, operate only in the first days of life, and their neural substrate actually atrophies when the work is done. Also, given that some of the constraints built into core knowledge representations are overturned in the course of explicit theory building, it is at least possible that core cognition systems themselves might be overridden in the course of development.
>
> (Carey, 2011b, p. 117)

Her commitment to the view that at least some dedicated input analyzers are innate dovetails with the definition of innateness as developmental essentiality that we introduced above. Consequently, the view that some concepts are innate because they are embedded in innate metaphysical transducers avoids the difficulty of accounting for how the rich and complex conceptual repertoire of most adults' minds can be built out of innate concepts. On this view, the innate conceptual resources are needed only to get learning started, and not to provide the ingredients for all concepts learned over the whole of cognitive development.[3] Whether or not these resources persist, and if so for how long, are empirical problems left open by the definition of innateness as developmental essentiality and which remain, so far as we know, unresolved.

But, as a scientific matter only, Carey could be wrong. It could be that further research yields compelling reasons to be skeptical of the existence of a suite of innate, dedicated input analyzers. Nevertheless, were that to occur, there would still be good metaphysical reasons to remain committed to the existence of innate transducers which mediate information transfer between the cognitive system and the other systems in BBE networks.

That said, this line of reasoning does not touch on a deeper outstanding problem: Why should anyone posit cognition at all? Metaphysical transducers are necessary *only if* we must explain how the cognitive system plays a functional role within a larger system of systems forming a BBE network. If cognition is not real, then there is no reason to posit innate cognitive transducers that come equipped with at least a handful of endogenous conceptual representations.

6. Consilience and the Choice between the BBE and the CBBE View

So, with that, we now arrive at the argument for preferring our extended view over Smith et al.'s restricted view. The specific question here is whether a cognitive system should be posited along with the systems (of systems) constituting the body, the brain, and the environment. We believe that you should posit CBBE networks instead of only BBE networks because doing so allows you to explain more scientific data than would otherwise be possible.

What kind of scientific data can only be explained by the extended CBBE view? This is any data that fits David Marr's operational definition of psychological

computation: A psychological process is computational if the process can be "characterized as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated" (Marr, 2010, p. 24). Note that Marr's definition refers to two logically distinct properties, properties that are jointly necessary in order to establish evidence of computational processes. The first is evidence of some kind of mapping between sets of information, such that the latter set can be treated as some kind of (possibly amplitative) transformation of the former set. The second is some kind of evidence that the relevant transformation is normatively appropriate for the situation or context: In some non-arbitrary sense, it is one a mind *should do*. Put another way, then, empirical evidence of psychological computation just is evidence of the rational processing of information (cf., Rumelhart & McClelland, 1985).

And there is a very large amount of exactly that kind of evidence. For example, consider two decades' worth of experiments that, taken together, demonstrate that both children and adults make inferences that seem to reflect unconscious knowledge of certain basic principles of logic and probability (Xu, 2007; Xu & Tenenbaum, 2007 Buchsbaum et al., 2011; Denison & Xu, 2012; Xu & Kushnir, 2012; Xu & Kushnir, 2013; Gopnik & Bonawitz, 2015; Wellman et al., 2016). Indeed, by about the age of 4, children have the ability to recognize when information is relevant (Southgate et al., 2009), when information is supportive of generalizations (Sim & Xu, 2017), when information is evidence of causation (Gopnik et al., 2004; Sim et al., 2017), and when information can be expressed on an ordinal scale (Hu et al., 2015). Of course, the mind's sensitivity to these different kinds and uses of information is not neutral: Frequently, information is used as evidence—that is, the information is used to drive changes in belief or motivation, changes that are themselves consistent with certain deep principles of rationality (Xu, 2007; Xu, 2011; Fedyk & Xu, 2017). This information is used, that is to say, in roughly the way it *should* be used *if* it is to be used *rationally*, satisfying Marr's operational definition of computational processing.

This evidence demonstrates that there is a meaningful scientific choice between a theory of cognitive development that posits only components of BBE networks and a theory of cognitive development that posits, in addition, the existence of a *sui generis* cognitive system. Considerations of scientific consilience (cf., Wilson, 1999; Cantor et al., 2018) favor the latter theory of cognitive development—since only a theory which posits a cognitive system that is a computational system is able to explain *both* the impressive amount of empirical data that Smith et al. survey in their chapter, *and* the scientific data that is evidence of computational processing. More scientific data can be explained by our extended view of the metaphysics of cognitive development than Smith et al.'s more restricted ontology.

7. Conclusion

The picture of cognitive architecture that we want to endorse is, at bottom, this: There is a set of designated input analyzers that are innate to the cognitive system

itself, plus a central reasoning system that scientists can study and understand by relying upon principles of rationality. The system as a whole is stable because the brain's innate degeneracy acts as one causal buffer—and not the only causal buffer—for the cognitive mind. The cognitive system is co-instantiated with the brain's dynamic networks—and, in this way, it is no different than all other real-world computational systems, given that (non-zero-error) computational systems are always, and can only be, co-instantiated with physical systems.

So, where is the mind? It is somewhere between the ears and behind the eyes, because it is co-instantiated with the brain. However, if we are instead asking about the functional location of the mind, then we can now be slightly more precise. The mind's functional location is given by asking how embedding a computational system within a BBE network extends the functional capabilities of the network. The most consequential of these increases seems to be to allow the resulting CBBE network to realize patterns of normative thought, which thereby dramatically amplify the range of context-appropriate behaviors available to the network. Or, put more simply, a cognitive system confers *rationality*—the capacity for different kinds of (epistemic, statistical, logical, moral, practical) principles to influence thought and regulate behavior. This dramatically increases the range of learning that is possible for our species (Tomasello, 2014), but it also dramatically deepens the sources of error, confusion, and mistakes as well. It is only by having a mind, after all, that someone can seem to discover reasons to doubt the existence of the same.

Notes

1 That said, Smith et al. are correct that there is no established definition of "innate"—for different examples see Kitcher (2001), Griffiths & Machery (2008), Griffiths (2002), and Ariew (1996). Of course, this shows neither that innateness is not real nor that the various definitions are confused or incoherent. In fact, we should expect a small family of potentially incommensurable concepts for some kind to develop as a byproduct of routine scientific investigation into the kind. Clusters of incommensurable concepts may sometimes be signs of inductive progress; we are, therefore, happy to add another concept to the cluster.

2 See also Bullmore & Sporns (2012) and Dehaene & Changeux (2011) for richer analyses of how different mental functions may stand in a many-to-one relationship with various forms and instantiations of neuronal connectivity. See also Aizawa (2015) for discussion of several complementary philosophical issues.

3 It is also helpful to point out that dedicated input analyzers and their innate conceptual resources are not necessarily Fodorian modules. Fodorian modules are a type of non-cognitive to cognitive metaphysical transducer, but they are not the only possible transducer which can perform that function. To see this, consider how Fodor describes a hypothetical module:

> A parser for [a language] L contains a grammar of L. What it does when it does its thing is, it infers from certain acoustic properties of a token to a characterization of certain of the distal causes of the token (e.g., to the speaker's intention that the utterance should be a token of a certain linguistic type). Premises of this inference can include whatever information about the acoustics of the token the

> mechanisms of sensory transduction provide, whatever information about the linguistic types in L the internally represented grammar provides, and nothing else.
> (Fodor, 1984, p. 37)

Separately, Fodor discusses cognitive transducers (Fodor, 1987). Furthermore, note that transduction is mentioned by Fodor, but it refers to processing prior to the module. But this language parser, too, is a metaphysical transducer: It converts acoustic information into lexical (or *lexicalizable*) information. It is an example, thus, of a transducer operating on the output of a transducer. So, again, Fodorian modules are a kind of metaphysical transducer, but they are not the only kind.

References

Aizawa, K. (2015). What is this cognition that is supposed to be embodied? *Philosophical Psychology*, 28(6), 755–75.

Ariew, A. (1996). Innateness and canalization. *Philosophy of Science*, 63, S19–S27.

Baillargeon, R., et al. (2012). Object individuation and physical reasoning in infancy: An integrative account. *Language Learning and Development: The Official Journal of the Society for Language Development*, 8(1), 4–46.

Bermudez, J. L. (2007). *Philosophy of Psychology: Contemporary Readings*. London: Routledge.

Boyd, R. N. (1993). Metaphor and theory change. In A. Ortony (Ed.), *Metaphor and Thought*, second edition. Cambridge: Cambridge University Press.

Buchsbaum, D. et al. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, 120(3), 331–40.

Bullmore, E., & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews: Neuroscience*, 13(5), 336–49.

Cantor, P. et al. (2018). Malleability, plasticity, and individuality: How children learn and develop in context. *Applied Developmental Science*, 23, 1–31.

Carey, S. (2011a). *The Origin of Concepts*, reprint edition. Oxford: Oxford University Press.

Carey, S. (2011b). Precis of "The Origin of Concepts." *Behavioral and Brain Sciences*, 34(3), 113–24.

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, 68(1), 53–74.

Craver, C. F. & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology & Philosophy*, 22(4), 547–63.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. A. Wilson (Eds.), *Explanation and Cognition*, 117–45. Cambridge, MA: MIT Press.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 4. http://www.jstor.org/stable/3129898

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–27.

Denison, S., & Xu, F. (2012). Probabilistic inference in human infants. *Advances in Child Development and Behavior*, 43, 27–58.

Edelman, G. M. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24), 13763–8.

Eliasmith, C. (2007). How to build a brain: From function to implementation. *Synthese*, 159(3), 373–88.

Eliasmith, C., & Anderson, C. H. (2004). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Fedyk, M., & Xu, F. (2018). The epistemology of rational constructivism. *Review of Philosophy and Psychology*, 9(2), 343–62. https://doi.org/10.1007/s13164-017-0372-1

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2), 97–115.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Fodor, J. A. (1984). Observation reconsidered. *Philosophy of Science*, 51(1), 23–43.

Fodor, J. A. (1987). Why paramecia don't have mental representations. *Midwest Studies in Philosophy*. http://onlinelibrary.wiley.com/doi/10.1111/j.1475–4975.1987.tb00532.x/full

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Noûs*, 31, 149–63.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford: Oxford University Press.

Gopnik, A. et al. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 3–32.

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2), 75–86.

Griffiths, P. E. (2002). What is innateness? *Monist*, 85(1), 70–85.

Griffiths, P. E., & Machery, E. (2008). Innateness, canalization, and "biologicizing the mind." *Philosophical Psychology*, 21(3), 397–414.

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Belknap Press, an imprint of Harvard University Press.

Hu, J. et al. (2015). Preschoolers' understanding of graded preferences. *Cognitive Development*, 36, 93–102.

Istrail, S., De-Leon, S. B.-T., & Davidson, E. H. (2007). The regulatory genome and the computer. *Developmental Biology*, 310(2), 187–95.

Jablonka, E., Lamb, M. J., & Zeligowski, A. (2014). *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*, revised edition. Cambridge, MA: MIT Press.

Kim, J. (1990). Supervenience as a philosophical concept. *Metaphilosophy*, 21(1–2), 1–27.

Kitcher, P. (2001). Battling the undead: How (and how not) to resist genetic determinism. *Thinking About Evolution: Historical, Philosophical, and Political Perspectives*, 2, 396–414.

Koutroufinis, S. A. (2017). Organism, machine, process: Towards a process ontology for organismic dynamics. *Organisms: Journal of Biological Sciences*, 1(1), 23–44.

Love, A. (2018). Developmental mechanisms. In S. Glennan & P. Illari (Eds.), *The Routledge Handbook of the Philosophy of Mechanisms*. New York: Routledge.

Marcus, G. F. (2006). Cognitive architecture and descent with modification. *Cognition*, 101(2), 443–65.

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Matthews, L. J., & Tabery, J. (2017). Mechanisms and the metaphysics of causation. *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, 115. London: Routledge.

McLaughlin, B., & Bennett, K. (2018). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2018/entries/supervenience/

Mizushima, N., & Levine, B. (2010). Autophagy in mammalian development and differentiation. *Nature Cell Biology*, 12(9), 823–30.

Packard, A. (2001). A "neural" net that can be seen with the naked eye. In W. Backhaus (Ed.), *Neuronal Coding of Perceptual Systems*, 397–402. Singapore: World Scientific Publishing Co.

Park, H.-J., & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science*, 342(6158), 1238411.

Păun, G., & Rozenberg, G. (2002). A guide to membrane computing. *Theoretical Computer Science*, 287(1), 73–100.

Reik, W., Dean, W., & Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, 293(5532), 1089–93.

Rumelhart, D. E., & McClelland, J. L. (1985). Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General*, 114(2), 193–7.

Samuels, R. (2000). Massively modular minds: Evolutionary psychology and cognitive architecture. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the Human Mind: Modularity, Language and Meta-cognition*, 13–46. Cambridge: Cambridge University Press.

Schulz, A. W. (2018). *Efficient Cognition: The Evolution of Representational Decision Making*. Cambridge, MA: MIT Press.

Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1), 132–50.

Sim, Z. L., Mahal, K. K. & Xu, F. (2017). Learning about causal systems through play. *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* https://mindmodeling.org/cogsci2017/papers/0210/paper0210.pdf

Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology*, 53(4), 642–51.

Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review: DR*, 25(3), 278–98.

Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: Cognitive Architecture*. Cambridge, MA: MIT Press.

Southgate, V., Chevallier, C., & Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Developmental Science*, 12(6), 1013–19.

Tabery, J. G. (2004). Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science*, 71(1), 1–15.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6), 3257–62.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind: A Quarterly Review of Psychology and Philosophy*, 59(236), 433–60.

Wang, S.-H., & Baillargeon, R. (2006). Infants' physical knowledge affects their change detection. *Developmental Science*, 9(2), 173–81.

Wellman, H. M. et al. (2016). Infants use statistical sampling to understand the psychological world. *Infancy: The Official Journal of the International Society on Infant Studies*, 21(5), 668–76.

Whittle, A. (2007). The co-instantiation thesis. *Australasian Journal of Philosophy*, 85(1), 61–79.

Wilson, E. O. (1999). *Consilience: The Unity of Knowledge*. New York: Vintage Books.

Xu, F. (2007). Rational statistical inference and cognitive development. *The Innate Mind: Foundations and the Future*, 3, 199–215.

Xu, F. (2011). Rational constructivism, statistical inference, and core cognition. *Behavioral and Brain Sciences*, 34(03), 151–2.

Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, 30(2), 111–53.

Xu, F., & Kushnir, T. (2012). *Rational Constructivism in Cognitive Development*. Cambridge, MA: Academic Press.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, 22(1), 28–32.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–72.

Further Readings for Part II

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.
Highly influential book uses work from developmental psychology to argue for the existence of innate concepts within core cognitive systems as well as the ability to acquire novel concepts via a process of Quinean bootstrapping.

Cowie, F. (1999). *What's Within? Nativism Reconsidered*. New York: Oxford University.
Draws on work from across cognitive science to critique arguments for nativist theses regarding language acquisition and innate concepts from Noam Chomsky and Jerry Fodor, respectively.

Fodor, J. A. (1981). The present status of the innateness controversy. In J. A. Fodor (Ed.), *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.
The strongest statement of Fodor's notorious argument for radical concept nativism.

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology*, 60, 115–40.
Draws on work in developmental psychology to argue that innate capacities, social input, and direct observation are all important influences on children's acquisition of concepts.

Gross, S., & Rey, G. (2012). Innateness. In E. Margolis, R. Samuels, & S. Stich (Eds.), *Oxford Handbook of Philosophy of Cognitive Science*, 318–60. New York: Oxford University Press.
A lengthy survey of the contemporary debate on what innateness is and whether concepts are innate.

Prinz, J. J. (2004). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
Draws on philosophical and empirical literature to argue for a new form of concept empiricism and against nativist theses from Chomsky and Fodor.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1), 89–96.
Draws on research on non-human primates as well as human infants, children, and adults to argue for the existence of domain-specific systems of core cognition that represent objects, actions, numbers, places, and potentially social partners.

Study Questions for Part II

1) According to Smith and colleagues, what is the brain–behavior–environment network, and how can it explain human behavior without positing the existence of concepts?
2) According to Smith and colleagues, how have recent advances rendered moot traditional questions about innateness?
3) According to Smith and colleagues, which questions should we investigate in place of questions about the origins of concepts?
4) According to Fedyk and Xu, which concepts are innate, and what makes them innate?
5) According to Fedyk and Xu, how does positing innate concepts help explain phenomena that could not otherwise be adequately explained?
6) Why do Fedyk and Xu disagree with Smith and colleagues about the existence of a *sui generis* cognitive system?

Part III

What Is the Role of the Body in Cognition?



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5 Embodied Cognition and the Neural Reuse Hypothesis

Julian Kiverstein

1. Introduction and Overview

At the heart of the embodied cognition research program is a hypothesis about the evolution of human "higher" cognitive processes. The hypothesis is one of *strong functional continuity* of phylogenetically older functions tied to perception and action and evolutionarily more recent human cognitive achievements that differentiate humans from other animals, such as language, mathematics, reasoning, and theory of mind. These distinctively human cognitive capacities are conceptualized not as species-specific cognitive adaptations. They are hypothesized to be instead the evolutionary outcome of redeploying and repurposing highly developed perception–action capacities in the service of new ends (Barrett, 2011; Anderson, 2014, 2015; Rietveld & Kiverstein, 2014). Support for this evolutionary hypothesis comes in part from a wide range of findings in cognitive neuroscience that show how regions of the human brain that were once thought to have specialized sensorimotor functions are also activated in many high-level cognitive tasks.[1] Additional support in cognitive neuroscience comes from findings that show the activation of classical emotion areas such as the amygdala and anterior insula in a wide range of cognitive tasks. These findings suggest that cognitive processes such as attention and executive function make extensive use of emotional evaluations of the world whose evolutionary origins are arguably inextricably tied up with perception and action.[2]

In what follows I will focus on the first type of findings which I will group under the heading of neural reuse. I will be concerned with two different interpretations of neural reuse that form the basis for two very different theories of what embodied cognition is. The first interpretation is found in recent work of Alvin Goldman who, drawing on Michael Anderson, has used the concept of neural reuse to argue for what he labels a "moderate" approach to embodied cognition. According to Goldman, there is nothing in the idea of embodiment that calls for a revision in the beliefs of the cognitive science orthodoxy about how human cognition works. Goldman writes that his conception of embodied cognition is "fully in sync with existing empirical research and raises no questions, for example, about such staples of traditional cognitive science as mental representations or computational processing."[3] Cognition should still be thought of as fundamentally a matter of computation over inner mental representations by the brain. Embodied

cognition just happens to be a subclass of cognition so conceived in which the representations over which the brain's computational processes operate are ones with a special body format or code.

The second concept of embodied cognition I will outline comes from Michael Anderson himself and is somewhat more revisionary in its implications for traditional ideas about cognition in cognitive science. I will side with Anderson in what follows and argue that Goldman is mistaken in his verdict about what embodied cognition means for cognitive science. I will argue that embodied cognition should be thought of as thoroughly pragmatic and action-oriented. This is a theme in research on embodied cognition that is conspicuously absent in Goldman's treatment of embodied cognition.

The chapter proceeds as follows. First, I provide an overview of the two conceptions of embodied cognition just mentioned and sketch the debate in more detail. In section 3, I outline something of the evolutionary background to neural reuse and contrast the resulting theoretical perspective with more standard ideas in evolutionary psychology that take the mind to have a massively modular organization. In section 4, I return to the concept of neural reuse and explain how it relates to the evolutionary perspective set out in section 2. Section 5 ends by using what we have learned about neural reuse earlier in the chapter to raise some problems for Goldman.

2. Two Conceptions of Embodied Cognition

In recent work, Alvin Goldman has used the evidence for neural reuse as part of an argument for a *moderate* approach to embodied cognition (Goldman, 2012, 2014). Goldman takes cognition to be embodied because some (but crucially not all) cognitive processes make use of representational formats that are bodily in nature. Goldman describes representations in bodily formats ("B-formats") as representational systems "dedicated to representing bodily subject matters" (2014, p. 100). Goldman gives the following examples of representations in bodily formats: somatosensory systems that function to represent haptic information; interoceptive systems that represent pain, tickle, itch, temperature, muscular and visceral sensations; and motor systems that control motor behavior. Goldman takes cognition to be embodied when representational systems that are specialized for representing bodily subject matters, such as the ones just mentioned, are reused or redeployed in the performance of cognitive tasks such as language processing, theory of mind, prospective planning, and mathematical reasoning. Embodied cognition so understood remains primarily a neural phenomenon. Cognition turns out to count as embodied only when it makes use of neuronal representational systems whose evolutionary function concerns the body in some way.

Goldman's treatment of embodied cognition is consistent with a functional separation of perception from action which, I shall argue, is strongly challenged by embodied cognitive science. In traditional cognitive science one often finds distinctions made between central cognitive processes, often taken to be non-modular in their workings, and perception and action, which are taken to be modular in their processing. An embodied approach to cognitive science begins

with a rejection of this separation of cognition from perception and action (Clark, 1997; Hurley, 1998; Brooks, 1999; Anderson, 2003, 2014; Cisek & Kalaska, 2010; Barrett, 2011). Goldman seems to also share this rejection. We will eventually see, however, that his moderate conception of embodied cognition leaves in place a problematic separation of perception from action. Indeed, Goldman's "moderate" account of embodied cognition is consistent with a more classical, modular understanding of perception and action, and a brain-centered theory of cognitive processes which I shall argue is challenged by the embodied cognition research program.

An alternative perspective on embodied cognition that also takes its lead from the phenomenon of neural reuse can be found in the recent work of Michael Anderson (Anderson, 2010, 2014). Anderson offers a theoretical perspective on brain evolution according to which brains in general evolved for the control of the exploratory activities of environmentally situated organisms. The brain controls behavior so as to bring about changes in the organism's relation to the environment that are desired, and that fit with the organism's needs. What the organism perceives is the value of its current relation to the environment. Anderson follows the American pragmatist philosopher John Dewey in arguing that there is no separating perception from action because every perception is accompanied and preceded by some action (Dewey, 1896; Anderson, 2014, ch. 5). He argues that these finely honed capacities for skillfully interacting with the environment are also partly constitutive of distinctively human capacities for thinking. Our sensorimotor capacities for interacting with the environment are redeployed in social and cultural contexts, for example in the coordination of our actions with other people which is the basis for our communicative abilities (Anderson, 2014, ch. 7).[4] Anderson's vision of embodied cognition is more all-encompassing than Goldman's. A science of embodied cognition does not only cover cognitive processes that make use of bodily formatted representations. It aims to explain how the entire spectrum of distinctively human capacities for higher-order cognition could have evolved through the reuse and repurposing of more basic sensorimotor capacities.

Goldman's moderate approach to embodied cognition suggests that little needs to change with respect to the methodologies of the cognitive sciences. He recommends that cognitive scientists investigate both the primary use of bodily formatted representations and the reuse of these representational systems in other cognitive domains. Goldman rightly points out, however, that "this is not a terribly revolutionary proclamation. Substantial sectors of cognitive science are already doing this" (Goldman, 2012, p. 81). If Goldman's proposal is along the right lines, there is nothing in the recent flurry of interest in embodied cognition that calls for a rejection or replacement of the explanatory tools of traditional cognitive science. Cognitive scientists can continue to work with a concept of cognition as the process by which the brain builds and computes over inner mental representations.

Anderson's analysis of embodied cognition allows by contrast that many cognitive processes don't only take place in the brain, but in the sensorimotor dynamic coupling of the agent with its ecological setting. It is the body itself in its interactions with a richly resourceful environment that is involved in the

performance of cognitive tasks, over and above any representations of the body that might be involved in the performance of such tasks.[5] Wilson and Gablonka (2013, p. 1) capture the core idea when they write:

> the brain is not the sole cognitive resource we have available to us to solve problems. Our bodies and their perceptually guided motions through the world do much of the work required to meet our goals replacing the need for complex internal representations.

The organism uses a mix of neural, bodily, and environmental resources, assembled on the fly through the dynamic coupling of the organism and the environment, and coordinated and controlled in ways that are specific to the task at hand (Runeson, 1977; Bingham, 1988). Some of the work that cognitive scientists might have ordinarily supposed is done by computing over internal mental representations and stored knowledge is instead done by acting on the environment. One of the important lessons of research on embodied cognition is thus that cognitive scientists shouldn't make appeal to more complex internal computational solutions to a problem until they have ruled out the possibility of simpler ecological solutions. Often these solutions will depend on the coordination and control of task-specific resources distributed over the brain, body, and environment. However, explaining cognitive systems that span the organism–environment boundary requires new explanatory tools for capturing the dynamics of systems whose defining parameters can lie on either side of the organism's skin. A science of embodied cognition may therefore require a radical overhaul of the explanatory tools of traditional cognitive science. It may even call for a replacement of the traditional theoretical framework which conceives of cognitive states and processes as fundamentally computational and representational in their workings.[6]

3. Neural Reuse: An Evolutionary Scenario

Humans are evolved creatures. It is therefore natural to think that our nature as evolved beings might have something to tell us about the psychological capacities humans possess today. How might evolutionary ideas contribute to the scientific study of the mind? My aim in this section will be to outline an evolutionary hypothesis I find in research on embodied cognition, and which makes the best sense of the phenomenon of neural reuse. This evolutionary hypothesis will also eventually enable us to see the shortcomings of Goldman's account of the embodiment of cognitive processes.

A familiar use of evolutionary theory within psychology has been to argue for a theory of the cognitive architecture of the human mind as massively modular (Sperber, 1994; Cosmides & Tooby, 1997, 2013; Carruthers, 2006). Each module is a functionally specialized mechanism dedicated to processing information relating to a specific domain. The massive modularity hypothesis (MMH) is based on an assumption that human psychological capacities are computational mechanisms that contributed to our hunter-gatherer ancestors' adaptedness to the environment—the so-called "Environment of Evolutionary Adaptedness" (EEA);

see Bowlby (1969) and Symons (1979). It is assumed that each module is an algorithmic mechanism that is functionally dedicated to solving a specific problem repeatedly encountered in the EEA. Organisms no doubt faced many problems and thus would have needed many specialized computational mechanisms for solving these problems. Evolved psychological traits are thus computational mechanisms built to take in specialized information, and to process this information so as to deliver solutions to specific adaptive problems. The MMH thus moves from the form of a problem our ancestors probably encountered in their natural environments, to the algorithmic function a given module would most likely need to have executed if it were to have contributed to an organism's fitness.[7]

The MMH takes there to be an *asymmetric* relation between the organism and its environment. The environment exerts selection pressures on the organism in the form of adaptive problems. Psychological mechanisms are sculpted and molded by natural selection to process certain types of information in specialized ways so as to solve these adaptive problems. The resulting mechanisms are adapted to function in very specific domains. These specific domains are, however, conceived of as wholly independent from and external to the organism, repeatedly posing problems to the organism of which the evolved psychological mechanisms are solutions. As Richard Lewontin has noted in critiquing adaptationist theories of natural selection:

> In this view the organism is the object of evolutionary forces, the passive nexus of independent external and internal forces, one generating "problems" at random with respect to the organism, the other generating "solutions" at random with respect to the environment.
>
> (Lewontin, 2001, p. 47; quoted by Walsh, 2014, p. 218)

Embodied cognitive science is by contrast committed to an animal–environment mutualism that can be traced back to the American pragmatists. Dewey wrote for instance:

> The idea of environment is a necessity to the idea of organism, and with the conception of environment comes the impossibility of considering psychical life as an individual, isolated thing developing in a vacuum.
>
> (Dewey, 1896, p. 285)

This "mutualist" perspective on evolved psychological traits claims that the relation between an organism and its environment is *complementary* and *symmetrical*. The environment isn't a neutral space, indifferent to the living forms that inhabit it, into which animals are inserted. The environment of an organism is instead rich with *affordances*. The latter term originates with the ecological psychologist J. J. Gibson, an important precursor to, and a continuing inspiration for, work in embodied cognitive science (see e.g. Chemero, 2009). Gibson coined the term "affordance" to refer to what the environment "offers the animal, what it provides or furnishes, for good or ill" (Gibson, 1979, p. 127). The environment is thus the "condition of an animal's existence" and is not external to the animal (Walsh,

2012). It provides or furnishes opportunities that are "propitious" and which the animal exploits in ways that contribute to its flourishing. It offers challenges that are unpropitious which the organism seeks to avoid, or acts so as to change for the better (Walsh, 2014, p. 223).

Think of niche construction, for example, and the diverse ways in which animals alter their environment in building nests, webs, and burrows so as to deal better with the challenges the environment poses to their survival. More generally, the organism develops psychological and behavioral capacities that complement and are tightly coordinated with the affordances of its environment. The environment in turn offers opportunities and poses challenges only for an animal with specific bodily abilities (Chemero, 2003, 2009; Rietveld & Kiverstein, 2014). Organism and environment thus form a complementary pair. I would even go so far as to agree with the anthropologist Tim Ingold that the organism and its environment form a single "indivisible totality" (Ingold, 2011, p. 19).

This mutualist perspective complements work in evolutionary developmental biology that shows how an organism's development can be among the engines of evolutionary change. In the modern synthesis theory of evolution, organisms only develop those phenotypic traits they inherit from their parents. Changes made to the organism during development do not get passed on to their offspring. It is only genes as replicators that are copied from parent to offspring that allow for the transmission of phenotypic traits across generations. Recent work in evolutionary developmental biology, however, has begun to argue for a role for phenotypic plasticity in evolutionary change. Phenotypic plasticity refers to "the property of a given genotype to produce different phenotypes in response to distinct environmental conditions" (Pigliucci, 2001, p. 1; see also West-Eberhard, 2003). For example, butterflies can develop different wing colourings depending on the season in which they are born, thus allowing them to better blend into their respective environments (Hiyama et al., 2012). The mechanisms that are involved in the development of an organism can produce a wide range of structures under different environmental conditions.

West-Eberhard (2005) shows how adaptive responsiveness of an organism's phenotypic traits to its environment can over time cause changes in the genetic structure of a population. Indeed Denis Walsh has argued that the "purposive maintenance of viability" by the organism is "the principal cause of evolutionary change" (Walsh, 2012, p. 201). Thus the adaptations of the organism are not explained exclusively by the external properties of the environment and the selection pressures these external properties exert on a population. The organism, through adaptive changes in its phenotypic traits that take place in development, plays an active role in modifying its relationship to the environment in ways that impact its chances of survival and reproduction. The animal develops a repertoire of phenotypic traits that are coordinated with the opportunities and challenges offered by the environment. The organism isn't born equipped with domain-specific psychological adaptations, but is, in the words of the anthropologist Tim Ingold, "a singular locus of creative growth within a continually unfolding field of relationships" (2011, pp. 4–5). Its psychological capacities are best thought of not as domain-specific psychological adaptations but as skills that belong to "the

whole organic being (indissolubly mind and body) situated in a richly structured environment" (Ingold, 2011, p. 5).

The massive modularity hypothesis predicts that the human brain ought to be neatly decomposable into functionally specialized circuits, each of which can be damaged without this impacting on the performance of other specialized circuits.[8] For example, there are circuits specialized for speech perception that process only representations of sound waves, and circuits that process numerosity that only take as inputs representations of distinct objects (Barrett & Kurzban, 2006, p. 630). By contrast, the mutualist perspective I am recommending would predict a neural organization that functions first and foremost in order to regulate the organism's interactions with an environment of affordances. As Michael Anderson nicely puts it, "the brain is best understood as first and foremost an action controller, responsible for managing the values of salient organism–environment relationships" (Anderson, 2014, p. xxii). The salient organism–environment relations are the affordances of the environment that are relevant for the organism because they are in some way propitious or unpropitious. Brain structures that were originally evolved for this purpose of controlling the organism's relation to the environment can be put to new uses, many of which relate to the affordances of the environment found more recently in human evolutionary history. For example, brain structures that originally served some sensorimotor function can be repurposed for language in humans that inhabit an environment rich with language (Evans & Levinson, 2009).

Embodied theories of cognition are thus committed to the following evolutionary hypothesis:

> The evolved nature of humans as perceiving and acting beings provides constraints on how human cognizing and thinking can be organised.

The roboticist Rodney Brooks captures the central idea well in a much quoted passage by fans of embodied cognition:

> problem solving behaviour, language, expert knowledge and application, and reason, are all pretty simple once the essence of being and reacting are available. That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction. This part of intelligence is where evolution has concentrated its time—it is much harder.
>
> (Brooks, 1999, p. 81)

In the next section I will show how evidence of neural reuse is best understood in the context of the evolutionary hypothesis that sensation and action constrain how higher-cognition and thinking is organized. The mutualist perspective on evolution I have been recommending predicts that evolutionarily newer psychological capacities are likely to have emerged in development by combining existing resources (both neural and environmental) in novel ways and for new ends. It follows that we ought to find brain circuits that respond to inputs from multiple

task domains. Brain regions and networks that originally evolved to function in perception–action domains will have later been exapted for functioning in different and unrelated task domains. The same brain region or circuit can thus be reused to process information relating to a variety of different and unrelated task domains.

4. Neural Reuse and Embodied Cognition

The substantial and growing evidence of neural reuse suggests that brains are not organized into systems that are domain-specific in the sense of responding only to inputs whose formal properties the circuits were designed to operate on (Barrett & Kurzban, 2006). For instance, Anderson and Pessoa (2011) examined the selective activation of 78 standard regions of the brain in over a thousand distinct experimental tasks, covering 11 distinct task domains. They found that each region was "active in an average of 95 tasks spread across nine cognitive domains" (Anderson, 2014, p. 10). Anderson concludes from this and other large-scale meta-analytic studies that "local neural structures are not highly selective and typically contribute to multiple tasks across domain boundaries" (ibid.). In other words, the brain circuits he studied do not seem to be domain-specific, but can instead contribute to processing in multiple distinct task domains.

Further evidence of neural reuse comes from behavioral studies in experimental psychology which demonstrate interference and competition effects. If the same brain circuits make functional contributions in different task domains, one would predict interference and competition when subjects are asked to perform competing cognitive tasks that depend on the same circuits. This is indeed what we find. Glenberg and Kaschak (2002) found, for instance, that when subjects are asked to make a response that is the opposite of that described in a sentence, they take longer to judge the meaningfulness of the sentence. For example, if they hear the sentence "close the drawer" and they must respond by pressing a button that requires a movement towards the body, they are slower to respond than when the response requires a movement away from the body. There is thus an interaction between the two conditions which implies "a shared component between these two different processes—movement and comprehension" (Anderson, 2014, p. 20).

Casasanto and Dykstra (2010) found a similar interaction effect for a memory recall task. They asked subjects to move marbles either up or down while recollecting experiences with either positive or negative valence. Participants retrieved memories faster when the direction of the movement was congruent with the valence of the memory (i.e., upward for positive memories and downward for negative memories). Subjects also retrieved more positive memories when making upwards movement than when making downward movements, and vice versa for negative memories.

The evidence for interaction effects like these is impressive, but it admits of a number of different explanations. Experiments such as these are often interpreted as supporting concept empiricism, the view that the vehicles of conceptual thought are grounded in perceptual and motor experience. Concept empiricists reject a view of conceptual thought as amodal, abstract symbolic representations

that bear an arbitrary relation to referents in the world. They argue instead for a view of conceptual thought as depending on a biological substrate that represents sensory and motor information. In conceptual thought we reactivate perceptual and motor areas in an offline, "as if", simulation mode.[9] Concept empiricism would predict the occurrence of neural reuse in domains in which people engage in conceptual thought. They would explain interference effects, like the ones I just described, in terms of a conflict between sensorimotor representations. The sensorimotor representations that are activated offline for the comprehension task (or the memory recall task) come into conflict with the sensorimotor representation activated online for the motor task because both the comprehension and motor task are making use of overlapping neural circuitry.

Interaction and competition effects are also naturally explained by so-called conceptual metaphor theories. These theories claim that understandings of entities gained from bodily experience and sensorimotor interaction with the world form the basis for what are called "image schemas." Image schemas represent, for instance, bodily experiences such as pushing and being pushed, moving objects, and experiencing forces. They are "imaginative, nonpropositional structures that organise experience at the level of bodily perceptions and movement" (Gibbs, 2006, p. 91; see also Johnson, 1987). Image schemas are metaphorically mapped onto more abstract domains of understanding and reasoning. For instance, we talk about "life as a journey" by using the SOURCE–PATH–GOAL image schema as a metaphor that captures something about the meaning of the concept "life" as it applies to persons. Lakoff and Gallese have applied the concept of neural reuse to explain the metaphorical mapping from image schema to abstract concepts. They have argued: "*The same neural substrate used in imagining is also used in understanding*" (Gallese & Lakoff, 2005, p. 456). They go on to propose that a "key aspect of human cognition is *neural exploitation*—the adaptation of sensory-motor brain mechanisms to serve new roles in reason and language, while retaining their original functions as well" (op. cit.).

While both conceptual metaphor theory and concept empiricism do indeed seem to fit very well with the phenomenon of neural reuse, it is important to recognize, as Michael Anderson has repeatedly stressed, that neural reuse is a much more general phenomenon. It goes well beyond the grounding of semantic meaning in bodily experience, covering many additional phenomena that are not so easily captured in these terms.[10] Neural reuse requires there to be some shared functional properties that are common to the sensorimotor domain in which the neural circuits originally functioned, and the evolutionarily more recent domain for which the circuits have been repurposed. Sometimes this functional relation will also support a semantic mapping or grounding, as highlighted by the two theories of meaning just discussed, but this won't always be the case.

I take concept empiricism and conceptual metaphor theory to be best understood as theories of what I shall call "grounded cognition."[11] They are primarily theories of meaning that share the claim that the meaning of seemingly abstract concepts is somehow grounded in bodily experiences. They differ in important respects in how they flesh out this thesis and in the role they assign to the body in grounding meaning. Embodied cognition, as I understand it, is a more general

theory that claims that higher-cognitive functions such as executive functions, language understanding, theory of mind, mathematical reasoning, and so on, reuse mechanisms that originally evolved for perception and action. Neural reuse I am proposing is best understood as providing evidence for this evolutionary hypothesis. What is the relation between grounded cognition and embodied cognition? I will follow Anderson in understanding embodied cognition to be the more general phenomenon. Grounded cognition may turn out to be a functional byproduct of this more general phenomenon of neural reuse.[12]

This evidence of widespread reuse of neural resources for multiple functions suggests that what parts of the brain do, may not line up neatly with any specialized cognitive functions. In earlier work, Anderson suggested that different brain regions may nevertheless carry out specific computational operations, and thus can be characterized as having specific "workings" consistent with many distinct high-level cognitive uses (Anderson, 2007, 2010). In his recent book, Anderson is more thorough-going in his rejection of the decomposability of the brain into functionally specialized circuits or components. He now talks about different regions and circuits as having sets of causal properties that dispose them to a range of different uses. Which properties are relevant to determining the precise functional contribution of a region will depend upon the partnerships and coalitions the region enters into. The component parts of brain networks are not functionally specialized components, but what Anderson calls Transiently Assembled Local Neural Subsystems (TALoNS). The functional properties of the region that manifest on any given occasion will be determined by the network of which the region is a part. The region can change which of its functional properties it instantiates based on coalitions and partnerships it forms with other regions (Anderson, 2014, p. 93).

Anderson calls the functional properties of a given TALoNS neuroscientifically relevant psychological (NRP) factors. NRP factors are the "primitive psychological factors or 'ingredients' that capture the underlying contributions of regions and networks to overall behaviour" (Anderson, 2014, p. 129). Anderson argues that the NRP factors all relate in some way to the brain's management of the organism's interactions with its environment.

Consider in this light work reported in Cisek and Kalaska (2010) on neurons in the posterior parietal cortex (PPC). There is a long-standing debate about whether these neurons have a perceptual, motor, or executive function. Cisek and Kalaska report studies that suggest the activation of the PPC "reflects all categories at once without respecting these theoretical distinctions" (2010, p. 274). Neurons in the PPC do not have dedicated perceptual or motor functions. They certainly are involved in directing visual attention to salient regions of the environment, but they are also active when making reaching movements or saccades in specific directions. Activity in the PPC is also modulated by variables associated with decision making such as utility and desirability, as well as in estimating the likelihood of an outcome (Colby & Duhamel, 1996; Colby & Goldberg, 1999; Anderson & Buneo, 2003).

Cisek and Kalaska (2010) make sense of findings like these by hypothesizing that the brain processes sensory information so as to ready the organism to respond to

multiple affordances the environment offers. Each bodily state of action-readiness is a response to something that is salient in the environment. These multiple relevant affordances then compete with each other as the brain gathers evidence for selecting between them (op. cit., p. 277). Action selection or decision making thus doesn't occur before action-specification or movement planning, but the two processes occur simultaneously in parallel. Factors relating to reward, costs, and risks can all interact with the selection process, biasing the competition between relevant invitations to action from the affordances of the environment in a particular direction. Cisek and Kalaska conclude their paper by explicitly making the connection to embodied cognition: "the neural systems that mediate the sensorimotor behaviour of our ancient ancestors may have provided the foundations for modern cognitive abilities, and their consideration may shed light on the neural mechanisms that underlie human thought" (p. 289).

NRP factors probably will not reflect existing psychological categories, and may not map onto traditional distinctions between perception, higher-cognition, and action. The psychological categories we employ in folk psychology may not map neatly onto brain organization (see also Feldman Barrett, 2009). In line with the mutualist perspective outlined in section 2, I am proposing that the NRP factors are likely to relate to the animal's practical activities in its ecological niche rich with affordances. Thus if we ask ourselves what explains the brain's patterns of response to stimuli, it is likely that what the brain is interested in foremost, are aspects of the animal's ecology that are relevant because of what they offer the animal. If we as cognitive neuroscientists are interested in the functional contribution that a pattern of brain responses makes to behavior, we should think in terms of how this pattern of activity might contribute to preparing the animal to act on multiple action possibilities, while at the same time allowing it to flexibly switch from one activity to another should the circumstances so require. The functional specialization we find in the brain is thus best understood in an ecological context. The neural correlates of decision making (an example of a "higher" executive cognitive function) are likely to be widely distributed in the brain and to be constituted by the same neural processes as are involved in action specification and online specification of movement.

So far I have been concentrating on sensorimotor behavior as my example of how to think about NRP factors. This is because I am thinking of the brain as a control system that is fundamentally in the business of regulating the organism's relation to the environment so as to improve the organism's grip in its skillful engagement with the affordances of the environment (Bruineberg & Rietveld, 2014; Kiverstein & Miller, 2015). The idea I want to tentatively explore next is that other higher-cognitive functions might have as their neural constituents TALoNS widely distributed across the brain. Each TALoNS has an NRP factor that relates to sensorimotor interaction and coordination with the affordances of the environment. However, TALoNS can be combined in novel ways to support the kinds of higher-cognitive capacities that are distinctive of human cognition.

Thus consider in the light of this idea the evidence that verbs such as "lick," "pick," and "kick" activate regions in the motor cortex involved in mouth movements, hand movements, and leg movements (Pulvermüller, 2005). Similarly,

Broca's area, long associated with language processing, has also been shown to be active in action and imagery-related tasks. Finally, the mirror neuron system located in the premotor cortex is known to also play an active role in social interaction and language processing. These findings suggest that brain regions whose NRPs relate to action and social interaction are being reused to also play a role in communication-related tasks (Müller & Basho, 2004).[13]

The function TALoNS are serving in communicative contexts may not be different in kind from the functions they originally served in regulating the agent's interactions with the environment. The environment humans inhabit is one that is rich with cultural and social practices. Many of the affordances of the human environment are constructed by us, and their perception depends on skills learned from other members of a social and cultural practice. The view I am sketching but cannot argue for here suggests that language might also be best thought of as a social and cultural practice. Language doesn't require its own specialized package of computational resources, as has long been argued by linguists working in the Cartesian tradition (Chomsky, 1956). The neural resources we need for acquiring a language may instead have evolved from redeployment of neural structures that play a control function in action and social interaction. The capacities that we associate with symbolic thought, such as the capacity to put together representations with a recursive structure, may thus be the result of combining a brain that functions to control action with complex sign systems that find their home in the social and cultural environments humans inhabit.

Language itself can naturally be thought of as a tool that humans have invented to coordinate their social interactions (Anderson, 2014, chapter 7). As such language is an augmentation or enhancement of the functions of the brain. Language performs what is effectively the same function as that of the brain of managing and regulating people's interactions with the environment, and with each other. However, it does so through the medium of syntactically and logically structured representations very different from anything found inside of the heads of language users. The transition from lower to higher cognition that we find in humans is thus not necessarily the result of having a brain wired to engage in symbolic thinking. It could instead turn out that it is the way humans are able to flexibly and reliably couple with an environment rich in social and cultural affordances that is explanatory of many of the highest achievements of human cognition.

5. On Goldman's Definition of Embodied Cognition

Alvin Goldman has proposed an account of the relation between neural reuse and embodied cognition that is very different from the one I have been outlining in this chapter. We have seen earlier how he argues that embodied cognition is "body-coded cognition" (Goldman, 2012, p. 72). It is a type of cognition in which bodily formatted representations are put to work for novel cognitive functions in domains that are distinct from those in which they originally functioned. What

is distinctive about research on embodied cognition, according to Goldman, is that it shows that bodily formatted representations show up in surprisingly many different cognitive domains, many more than cognitive scientists might have hitherto supposed.

Goldman compares the idea of neural reuse with his earlier work on simulation theories of mindreading, according to which subjects make use of their own decision-making systems in an offline simulation mode to arrive at predictions and explanations about the decision-making processes in other people (Goldman, 2006). I want to briefly pause to outline this simulation-based account of neural reuse more deeply. It forms the basis for an account of the relation between lower and higher cognition that differs from the one I've been recommending, and we will see how it also points to wider differences over the right way to think about embodied cognition.

Goldman's idea that neural reuse might naturally be thought of as the basis for mental simulation was an idea also explored by Susan Hurley in her later work on social cognition (e.g., Hurley, 2008a, 2008b). Hurley suggested that simulation can be thought of in terms of an intrapersonal similarity between the processes that a subject uses to make decisions and to form intentions, and the processes the subject uses to understand the decisions and intentions of others (Hurley, 2008a, 2008b). This intrapersonal similarity is reflected in the brain in the reuse of decision-making processes in a pretend, "as-if" mode to understand others. Gallese and Sinigaglia (2011) have recently made a similar proposal for the lower-level processing involved in motor control.[14] They suggest for instance:

> the activation of parieto-premotor cortical networks, which typically serve the purpose of representing and accomplishing a single motor goal (such as grasping something) or a hierarchy of motor goals (such as grasping something for bringing it to the mouth or for placing it at a specific location) might also serve the purpose of attributing that motor goal or motor intention to others.
>
> (p. 513)

Gallese and Sinigaglia claim that the representations of one's own and of other's actions share a common bodily format, which they conceive of along similar lines to Goldman. They understand this bodily format in terms of bodily constraints deriving from such factors as the biomechanical, dynamical, and postural properties of the body that constrain what can be represented. These constraints determine how a goal, or hierarchy of goals, are represented. The very same constraints apply also to the representations of the goal-directed behavior of other people. Gallese and Sinigaglia hypothesize that I understand the actions of another by reusing the motor representations I use when I myself act. Thus, action understanding is underwritten by neural reuse.[15] Goals are represented in a bodily format that doesn't distinguish between whether the goal is mine, or the goal of another agent. This is why my brain can readily reuse the same motor representations produced when I act to understand the actions of others.

Others have used the concept of simulation to propose explanations of planning and deliberation, as well as conceptual thinking, as we saw briefly above in my discussion of grounded theories of cognition. The idea, very roughly, is that the same sensorimotor circuits that are used in perceiving and acting are used in offline or simulation mode for planning (Grush, 2004; Pezzulo, 2008). It has long been known that visualizing and vision make use of overlapping processes in the brain. Thus when we visualize an object in order to perform mental rotation, for instance, the visual processes that would be active if we were to physically interact with the object are also active in an offline mode (Kosslyn et al., 2006). Similarly, when we plan an action, motor areas of the brain are active that would be active if we were actually performing the action (Jeannerod, 1994). Finally, when people think about a perceptual object such as a watermelon, they tend to activate perceptual representations associated with the object they are thinking about (Barsalou et al., 1999). In all of these examples, perceptual-action representations are being reused in offline or simulation mode for imagining, planning, and conceptual thinking.

Thus neural reuse may be taken to reflect an offline simulation-based mode of processing. Higher-cognitive processes reuse lower-cognitive processes with a perception-action function in a simulation mode. As Goldman understands embodied cognition, it will turn out that many examples of neural reuse are best understood as examples of simulation, but not of embodied cognition. This is because for Goldman there is only embodied cognition in those cases in which a body-formatted representation is used in a cognitive domain. There will be many examples of neural reuse that do not involve the reuse of body formatted representations. Thus, contrary to what I have been arguing so far, neural reuse and embodied cognition only partially overlap. The overlap is found in cases of simulation that depend on the processing of B-formatted representations.

This simulationist understanding of neural reuse (at least in Goldman's hands) leaves in place a separation of perception and action. This separation is challenged by the mutualist perspective on the evolution of human psychology I sketched above, in which what animals perceive is first and foremost the affordances of the environment. Consider in this light what Goldman has to say about the research by Dennis Proffitt and colleagues showing, for instance, that visual experiences of properties like size and distance are modulated by bodily factors such as fatigue, fitness, and other energetic states of the body. Thus in one well-known study, Bhalla and Proffitt found that the perceived steepness of the hill was affected by the weight of the backpack the subjects in the experiment were carrying (Bhalla & Proffitt, 1999).[16] Goldman rightly takes this study to be an example of embodied cognition. He argues that this study (and others I do not have space to discuss) show that representations of bodily states can determine how a visual property (e.g., the spatial incline of a hill) is represented. He interprets these findings as evidence that representations of bodily states are used by the visual system "to scale physical judgements about other (non-bodily) subject matters" (Goldman, 2012, p. 84). These representations of bodily states are characterized by Goldman as having a modulatory influence on visual processing. Thus the representations

of bodily states influence how the objective property of the visual property of steepness is represented. Goldman describes subjects as overestimating or underestimating the steepness of the incline, for instance. Subjects that were found to be physically fit, for example, judged the incline of the hill to be shallower than did the less fit counterparts. Fitness here is thus influencing the assessment by the visual system of the hill's angle of incline.

From an ecological perspective like the one I've been advancing in this paper, Bhalla and Proffitt's experiments are best interpreted not as investigating the visual perception of an objective quantity like the spatial incline of the hill. The experiments are better seen as investigating the perception of an affordance, the climbability of a hill. The perception of an affordance is, however, always relative to the abilities of the perceiver. Affordances are relational properties; they are what the environment offers to an animal with the necessary bodily capacities and abilities (Chemero, 2003; Rietveld & Kiverstein, 2014). Thus a hill is perceivable as being climbable only for a perceiver with the necessary locomotive abilities. In the Bhalla and Proffitt experiment, the different assessments of the hill's steepness are best understood not as estimations of the hill's incline but as tracking the ability of the subject to climb the hill, an ability that is partly hindered when carrying a heavy load. Moreover, once we accept that it is climbability that the subjects are perceiving, it is not so clear they are mistaken in their perception of the hill's incline when they under- or overestimate its steepness. Perhaps we should say instead that subjects are correctly assessing their ability to climb the hill, an ability that is affected by the weight of the backpack they are carrying, and by the other bodily factors the experimenters set out to investigate.

Goldman's characterization of embodied cognition relies on the idea that some brain systems are specialized for representing bodily matters. The examples he gives of brain systems that are functionally specialized for representing the body would no doubt be regarded as uncontroversial by most brain scientists. However, it is precisely the evidence for functional specialization in the brain that I've been arguing neural reuse requires us to question. We've seen above in section 4 how Anderson proposes the concept of TALoNS to capture the functional differentiation in the brain. Each local subsystem in the brain has causal dispositions that contribute to regulating, managing, and controlling the value of the current relationship of the organism with its environment. Local neural subsystems participate in multiple overlapping networks in ways that depend upon task and context. Each subsystem, however, does not have a fixed, stable, and static specialized function such as the representation of bodily states and conditions. They rather display functional biases that reflect their original function of managing the organism's interactions with the surrounding environment, or what I've described above in terms of the organism's responsiveness to the affordances of the environment.

What Goldman misses in his conception of B-formatted representations is the action-oriented or pragmatic nature of the subsystems that get reused in embodied cognition. He misses how the many subsystems in the brain concerned

with bodily subject matters function to prepare the organism for action in an environment of more or less inviting possibilities for action. Indeed, Goldman's conception of embodied cognitive processes is quite consistent with older ideas in cognitive science that took cognitive processes to be general-purpose computers that take input from perceptual modules and send commands back out to motoric modules. I have been arguing that such a separation of perception and action is wholly at odds with the embodied cognition research program. Action is better understood as for the control of perception: The brain produces the actions that produce the sensory inputs the organism desires. In perception we are responsive to the affordances of the environment, and only secondarily and occasionally—when it matters—to the objectively measurable properties of things. Nor are cognitive processes separable from perception and action since cognitive processes are constituted by reusing and combining in novel ways neural resources that serve a perception–action function.

6. Conclusion

Goldman's theory of embodied cognition is in the end a theory of *embrained* cognition. Indeed Shaun Gallagher has dubbed Goldman's approach to embodied cognition as "body snatching" (2015, p. 97)). Along with other body snatchers such as Rob Rupert (see e.g. Rupert, 2009), Goldman has devised

> a version of embodied cognition that leaves the body out of it. They still retain the term "embodied" but in fact, for them, the body, per se, is not necessarily involved in the real action of cognition. Rather, the real action, all the essential action, occurs in the brain. Indeed the body in this version of embodied cognition is "the body in the brain."
>
> (Gallagher, 2015, p. 3)

I've been arguing by contrast that the body in embodied cognition is not only in the brain, but the brain is rather always in a body that has a skillful grip on an environment of affordances. Embodied cognitive science does not mean business as usual for cognitive scientists. On the contrary, it means studying cognition as it takes place in the whole body and in the world. Cognition is embodied because cognitive processes are bound up with the life of the whole person or animal in their practical engagement with their surrounding environment.

Notes

1 For an excellent review of the evidence for neural reuse see part 1 of Anderson (2014).
2 See Pessoa (2013) for detailed arguments to this effect and Kiverstein and Miller (2015) for further discussion.
3 Quote from Goldman (2012), p. 72.
4 I've made similar arguments in a jointly authored paper currently under revision (Kiverstein & Roepstorff, unpublished manuscript). See also Kiverstein and Rietveld (2015).
5 owe this formulation in part to Goldman (2014, p. 93).

6 Explanations that focus on the dynamics of the organism–environment system as a whole have sometimes been taken to challenge the necessity of internal mental representations in the production of cognitive behaviors (Hutto & Myin, 2013; Chemero, 2009; Ramsey, 2007; Brooks, 1999). However, strictly speaking, these explanations are neutral on the role of representations, and in practice many proponents of embodied cognition have retained a somewhat weak commitment to a representationalist theory of the mind; see e.g. Clark (1997, 2015), Wheeler (2005), Anderson and Chemero (2009). I will also attempt to retain an attitude of neutrality on this controversial issue in what follows. For what it's worth, I tend to agree with Ramsey and Hutto and Myin that much of the empirical literature in neuroscience and connectionism that makes free use of the notion of representation could just as well be redescribed in non-representational terms. Difficult questions remain for anti-representationalists, however, about so-called "offline" cases of cognition, which trade in information about the distally absent or about counterfactual states of affairs. For a first stab at dealing with offline cognition in non-representational terms, see Kiverstein & Rietveld (2015) and also Degenaar & Myin (2014).

7 The MMH is standardly distinguished from a thesis of modest modularity as found in the writings of Jerry Fodor (Fodor, 1983; see also Samuels, 2006). Classical computational theories of mind took cognitive processes to function more or less like a general purpose digital computer equipped with various input and output modules that deliver perceptual information and produce motor behaviors. The thesis of massive modularity claims by contrast that the mind is composed entirely of modules. Goldman is, strictly speaking, neutral on the debate between massive and modest theories of modularity, though he acknowledges that his conception of bodily formatted representations, or representations that are specialised for processing specific types of information about the body, borrows from the literature on modularity (Goldman, 2014, p.100).

8 When defining modularity it is normal practice for philosophers to follow Fodor's list of features that are a rough guide to whether a system counts as modular. I will, however, follow Barrett and Kurzban's lead in defining "modularity" by reference to functional specialization (Barrett & Kurzban, 2006). They define the latter in terms of mechanisms that were selected for processing of "formally definable informational inputs" (op. cit., p. 630).

9 See e.g. Barsalou, (1999, 2009), Prinz (2002), Borghi et al. (2004), Pecher and Zwaan (2009), Jirak et al. (2010). Machery (2007) provides persuasive arguments that the evidence for concept empiricism is inconclusive. Proponents of an amodal theory of symbolic thought can allow, for instance, that often when concepts are tokened in thought there is accompanying or associated visual and motor imagery. It doesn't follow that this accompanying imagery plays a constituting role in conceptual thought. Moreover, there are reasons to doubt that the role of perceptual simulation generalises to all tasks whose performance is knowledge intensive.

10 I won't rehearse Anderson's examples here because of space constraints, but for further discussion see Anderson (2014, section 1.3) and Anderson (2010).

11 See also Kiverstein (2012).

12 For some discussion of this point see my commentary on Anderson (2010) in which I raise the possibility that reuse may sometimes depend on grounding (Kiverstein, 2010).

13 See Anderson (2014, section 7.1) for more discussion of this point.

14 Hurley acknowledges Vittorio Gallese as an influence on her thinking about mental simulation.

15 See also Hurley (2008b).

16 For an important critique of these findings see Firestone (2013) and Scholl & Firestone (2016). One of the issues these authors highlight is whether effects of bodily experiences on visual processing reflect post-perceptual processing, rather than a modulation of perceptual processing by the bodily abilities of the perceiver. To my mind Proffitt (2013) provides a convincing response to this worry but unfortunately I cannot engage with this debate in what remains of the chapter.

References

Anderson, M. L. 2003. Embodied cognition: A field guide. *Artificial Intelligence*, 149(1): 1–30.

Anderson, M. L. 2007. Massive redeployment, exaptation, and the functional integration of cognitive operations. *Synthese*, 159(3): 329–45.

Anderson, M. L. 2010. Neural reuse: A fundamental organisational principle of the brain. *Behavioural and Brain Sciences*, 33(4): 245–66.

Anderson, M. L. 2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, M. L. 2015. Precis of *After Phrenology. Behavioural and Brain Sciences*, 16: 1–22.

Anderson, M. L. & Pessoa, L. 2011. Quantifying the diversity of neural activations in individual regions. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2421–6) Austin, TX: Cognitive Science Society.

Anderson, R. A., & Buneo, C. A. 2003. Sensorimotor integration in posterior parietal cortex. *Advances in Neurology*, 93: 159–77.

Barrett, H. C., & Kurzban, R. 2006. Modularity in cognition: Framing the debate. *Psychological Review*, 113(3): 628–47.

Barrett, L. 2011. *Beyond the Brain: How Body and Environment Shape Animal and Human Minds.* Princeton, NJ: Princeton University Press

Barsalou, L. 1999. Perceptual symbol systems. *Behavioural and Brain Sciences*, 22: 577–660.

Barsalou, L. 2009. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521): 1281–1289.

Barsalou, L., Solomon, K. O., & Wu, L.-L. 1999. Perceptual simulation in conceptual tasks. In M. K. Hiraga, C. Sinha, & S. Wilcox (Eds.), *Cultural, Psychological, and Typological Issues in Cognitive Linguistics: Selected Papers of the Bi-Annual ICLA Meeting, Albuquerque, July 1995* (pp. 209–28). Amsterdam, NL: John Benjamins Publishing Company.

Bhalla, M., & Proffitt, D. 1999. Visual-motor recalibration in geographical slant perception. *Journal of Experimental Psychology: Human Perception and Performance*, 25: 1076–96.

Bingham, G. P. 1988. Task-specific devices and the perceptual bottleneck. *Human Movement Sciences*, 7: 225–64.

Borghi, A. M., Glenberg, A. M., & Kaschak, M. P. 2004. Putting words in perspective. *Memory and Cognition*, 32: 863–73.

Bowlby, J. 1969. *Attachment and Loss. Volume 1: Attachment*. New York, NY: Basic Books.

Brooks, R. 1999. Intelligence without representation. In his *Cambrian Intelligence: The Early History of the New AI* (pp.79–103). Cambridge, MA: MIT Press.

Bruineberg, J., & Rietveld, E. 2014. Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, 8: 599. doi: 10.3389/fnhum.2014.00599

Carruthers, P. 2006. *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Gloucestershire: Clarendon Press.

Casasanto, D., & Dijkstra, K. 2010. Motor action and emotional memory. *Cognition*, 115(1): 179–85.

Chemero, A. 2003. An outline of a theory of affordances. *Ecological Psychology*, 15(2): 181–95.
Chemero, A. 2009. *Radical Embodied Cognitive Science.* Cambridge, MA: MIT Press.
Chomsky, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3): 113–124.
Cisek, P., & Kalaska, J. F. 2010. Neural mechanisms for acting with a world full of action choices. *Annual Review of Neuroscience*, 33: 269–98.
Clark, A. 1997. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.
Clark, A. 2015. *Being There: Putting Brain, Body and World Together Again.* Cambridge, MA: MIT Press.
Colby, C. L., & Duhamel, J. R. 1996. Spatial representations for action in parietal cortex. *Brain Research: Cognitive Brain Research*, 5(1–2): 105–15.
Colby, C. L., & Goldberg, M. E. 1999. Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22(3): 19–49.
Cosmides, L., & Tooby, J. 1997/2013. Evolutionary psychology: A primer. Reprinted in E. Machery & S. Downes (Eds,), *Arguing About Human Nature: Contemporary Debates.* London, UK: Routledge Taylor Francis.
Degenaar, J., & Myin, E. 2014. Representation-hunger reconsidered. *Synthese*, 191: 3639–48.
Dewey, J. 1896. The reflex arc concept in psychology. *Psychological Review*, 3: 357–70.
Evans, N., & Levinson, S. C. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioural and Brain Sciences*, 32: 429–92.
Feldman Barrett, L. 2009. The future of psychology: Connecting mind to brain. *Perspectives on Psychological Science*, 4(4): 326–39.
Firestone, C. 2013. How "paternalistic" is spatial perception? Why wearing a heavy backpack doesn't – and couldn't – make hills look steeper. *Perspectives on Psychological Science*, 8(4): 455–73.
Firestone, C., & Scholl, B. J. 2016. Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, 39, 1–19.
Fodor, J. 1983. *The Modularity of Mind: An Essay on Faculty Psychology.* Cambridge, MA: MIT Press.
Gallagher, S. 2015. Invasion of the body snatchers: How embodied cognition is being disembodied. *Philosopher's Magazine*, April 2015: 96–102.
Gallese, V., & Lakoff, G. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22(3–4): 455–79.
Gallese, V., & Sinigaglia, C. 2011. What is so special about embodied simulation? *Trends in Cognitive Sciences*, 15(11), 512–19.
Gibbs, R. W. 2006. *Embodiment and Cognitive Science.* Cambridge, UK: Cambridge University Press.
Gibson, J. J. 1979. *The Ecological Approach to Visual Perception.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Glenberg, A. M., & Kaschak, M. P. 2002. Grounding language in action. *Psychonomic Bulletin and Review*, 9: 558–65.
Goldman, A. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford: Oxford University Press.
Goldman, A. 2012. A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, 3(1): 71–88.
Goldman, A. 2014. The bodily formats approach to embodied cognition. In U. Kriegel (Ed.), *Current Controversies in the Philosophy of Mind.* London, UK: Routledge.
Grush, R. 2004. The emulation theory of representation: Motor control, imagery and perception. *Behavioural and Brain Sciences*, 27: 377–442.

Hiyama, A., Taira, W., & Otaki, J. M. 2012. Colour pattern evolution in response to environmental stress in butterflies. *Frontiers in Genetics*, February 6, 2012.

Hurley, S. L. 1998. *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Hurley, S. L. 2008a. Understanding simulation. *Philosophy and Phenomenological Research*, 77(3): 755–74.

Hurley, S. L. 2008b. The shared circuits model: How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioural and Brain Sciences*, 31(1): 1–58.

Hutto, D., & Myin, E. 2013. *Radicalising Enactivism: Basic Minds without Content*. Cambridge, MA: MIT Press.

Ingold, T. 2011/2001. *The Perception of the Environment: Essays on Livelihood, Dwelling and Skill*. London, UK: Routledge, Taylor and Francis.

Jeannerod, M. 1994. The representing brain: Neural correlates of motor intention and imagery. *Behavioural and Brain Sciences,* 17(2): 187–202.

Jirak, D., Menz, M. M., Buccino, G., Borghi, A. M., & Binkofski, F. 2010. Grasping language: A short story on embodiment. *Consciousness and Cognition*, 19: 711–20.

Johnson, M. 1987. *The Body in the Mind*. Chicago, IL: University of Chicago Press.

Kiverstein, J. 2010. No bootstrapping without semantic inheritance: Commentary on Michael Anderson's Neural Reuse. *Behavioural and Brain Sciences*, 33: 279–80.

Kiverstein, J. 2012. The meaning of embodiment. *Topics in Cognitive Science*, 4(4): 740–58.

Kiverstein, J., & Miller, M. 2015. The embodied brain: Towards a radical embodied cognitive neuroscience. *Frontiers in Human Neuroscience*, 9(237): 1–12.

Kiverstein, J., & Rietveld, E. 2015. The primacy of skilled intentionality: On Hutto & Satne's *The Natural Origins of Content*. *Philosophia*, 43(3): 701–21.

Kiverstein, J., & Roepstorff, A. Unpublished manuscript. The proof of the pudding: Prediction, coordination and common ground.

Kosslyn, S., Thompson, W. L., & Ganis, G. 2006. *The Case for Mental Imagery.* Oxford, UK: Oxford University Press.

Lewontin, R. C. 2001. *The Triple Helix: Genes, Organisms and Environments.* Oxford, UK: Oxford University Press.

Machery, E. 2007. Concept empiricism: A methodological critique. *Cognition*, 104: 19–46.

Müller, R. A., & Basho, S. 2004. Are nonlinguistic functions in "Broca's area" prerequisites for language acquisition? fMRI findings from an ontogenetic viewpoint. *Brain and Language*, 89(2): 329–336.

Pecher, D., & Zwaan, R. A. 2009. *Grounding Cognition: The Role of Perception and Action in Memory, Language and Thinking.* Cambridge, UK: Cambridge University Press.

Pessoa, L. 2013. *The Cognitive-Emotional Brain: From Interaction to Integration*. Cambridge, MA: MIT Press.

Pezzulo, G. 2008. Coordinating with the future: The anticipatory nature of representation. *Minds and Machines*, 18: 179–225.

Pigliucci, M. 2001. *Phenotypic Plasticity: Beyond Nature and Nurture.* Baltimore, MD: Johns Hopkins University Press.

Prinz, J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis.* Cambridge, MA: MIT Press.

Proffitt, D. 2013. An embodied approach to perception: By what units are visual perceptions scaled? *Perspectives on Psychological Science,* 8(4): 474–83.

Ramsey, W. M. 2007. *Representation Reconsidered*. Cambridge, UK: Cambridge University Press.

Rietveld, E., & Kiverstein, J. 2014. A rich landscape of affordances. *Ecological Psychology*, 26(4): 325–52.

Runeson, S. 1977. On the possibility of "smart" perceptual mechanisms. *Scandinavian Journal of Psychology*, 18: 172–9.

Rupert, R. 2009. *Cognitive Systems and the Extended Mind*. New York, NY: Oxford University Press.

Samuels, R. 2006. Is the human mind massively modular? In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science.* London, UK: Blackwell, pp.37–56.

Sperber, D. 1994. The modularity of thought and the epidemiology of representation. In L. A. Hirshfield & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture.* Cambridge, UK: Cambridge University Press, pp. 39–67.

Symons, D. 1979. *The Evolution of Human Sexuality.* Oxford, UK: Oxford University Press.

Walsh, D. M. 2012. The struggle for life and the conditions of existence: Two interpretations of Darwinian Evolution. In F. Brinkworth & F. Weinert (Eds.), *Evolution 2.0: Implications of Darwinism in Philosophy and the Social and Natural Sciences.* Dordrecht, NL: Springer.

Walsh, D. M. 2014. The affordance landscape: The spatial metaphors of evolution. In G. Barker, E. Desjardins, & T. Pearce (Eds.), *Entangled Life: Organism and Environment in the Biological and Social Sciences.* Dordrecht, NL: Springer.

West-Eberhard, M. J. 2003. *Developmental Plasticity and Evolution*. Oxford, UK: Oxford University Press.

West-Eberhard, M. J. 2005. Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of the Sciences*, 102: 6543–9.

Wheeler, M. 2005. *Reconstructing the Cognitive World: The Next Step.* Cambridge, MA: MIT Press.

Wilson, A., & Golonka, S. 2013. Embodied cognition is not what you think it is. *Frontiers in Psychology*. http://dx.doi.org/10.3389/fpsyg.2013.00058

6 Rehashing Embodied Cognition and the Neural Reuse Hypothesis

Fred Adams

1. Introduction

I am pleased to be invited to comment on Kiverstein's chapter and on this important topic. The import of embodied approaches to cognition is that they challenge the received approach to cognition. That is, the traditional view of cognition is that perceptual systems of the body merely deliver information to central areas of the brain. In these central areas, information is converted into concepts and ideas and other forms of cognitive representation. Then after cognitive processing, reasoning, and planning to act, information is sent to the motor system, which does the mind's bidding. And so, as Susan Hurley (2001) so aptly put it, cognition is "sandwiched" between perception and action.

What is supposed to be new and exciting about embodied cognition is the claim that the sandwich view is false. Cognitive processing is supposed not to be confined to central areas of the brain. Rather, it is supposed to extend to perceptual areas, motor areas, and for some even to areas of the body beyond the brain. For those who accept "extended cognition," cognitive processing is even believed to extend beyond the boundaries of brain and body to world (Clark & Chalmers, 1998).

2. Kiverstein's Objections to Goldman

Within embodied cognition there is what I will call a weak view and a strong view. The weak view is that the body (including perceptual and motor areas of the brain) provides causal input to cognition but these inputs themselves do not constitute part of the cognitive processing. As an analogy, my circulatory system provides causal input to my cognitive system (in the form of nourishment), but circulation and nutrition do not constitute cognition. Similarly, the fuel injectors in my car's engine causally contribute to the engine's combustion by mixing air and gas so it can enter the cylinders, but injection does not constitute combustion. So there is a difference between mere causal contribution to cognition and cognition itself (Adams & Aizawa, 2008). This weak view of embodiment is what I called the received view or the traditional view of cognition above. On this weak view, activity in perceptual and motor areas is causally contributing to cognition but not itself constitutive of cognition.

The strong view of embodied cognition maintains that the inputs from perception and motor regions of the brain are not only providing causal inputs to cognition, but that processes in these brain regions themselves partially constitute cognition. It is the strong view of embodied cognition that excites most proponents of embodied cognition and it is this view that Kiverstein aims to defend in his chapter. Indeed, many proponents of the strong view maintain that cognition takes place in the body and in the interaction of brain, body, and world.

Kiverstein's objection to Goldman is that Goldman's interpretation of the "neural reuse" hypothesis conforms only to the weak view of embodied cognition, whereas Kiverstein wants to defend the stronger view. He wants the neural reuse hypothesis to imply that representations of the perceptual and motor systems partially constitute cognitive processes.

Here are Kiverstein's words:

> Indeed, Goldman's conception of embodied cognitive process is quite consistent with older ideas in cognitive science that took cognitive processes to be general-purpose computers that take input from perceptual modules and send commands back out to motoric modules. I have been arguing that such a separation of perception and action is wholly at odds with the embodied cognition research program. Action is better thought of as for the control of perception: the brain produces the actions that produce the sensory inputs the organism desires. In perception we are responsive to the affordances of the environment, and only secondarily and occasionally—when it matters—to the objectively measurable properties of things. Nor are cognitive processes separable from perception and action since cognitive processes are the result of reusing and combining in novel ways neural resources that serve a perception–action function.
>
> (Kiverstein, p. 102)

I take it that when Kiverstein says the perception and action processes are not "separable" from cognitive processes, he means not separable on the strong interpretation—that is, they partially constitute cognitive processes. One could also read this on the weak view as not "causally separable." And about that Goldman would agree. So the only way to read this quote as an objection to Goldman is by taking Kiverstein to be defending the strong view of embodied cognition and to be complaining because Goldman is only taking the weak view.

On Goldman's view, the "reuse" of representations involves taking bodily formatted (B-formatted) representations (perhaps of perception or motor activities) of the body and reusing them for cognitive purposes. But the representations exist within the central areas of the brain. They aren't located in perceptual areas or motor areas. The latter areas provide causal input to cognition via these B-formatted representations. But the activities in these areas do not themselves constitute cognitive processing. So the role of the body on Goldman's interpretation of "reuse" is causal not constitutive. Hence, Goldman's view of embodiment is the weak view, not the strong view of the role of the body in cognition. And it is this weak view that Kiverstein finds objectionable. It is not new. It is not a

threat to the more traditional way of thinking about cognition. It is not exciting (my words, not his).

Is it wrong for Goldman to take the weak view of embodiment? I'll discuss this below when considering matters from Goldman's perspective. It turns out that both Kiverstein and Larry Shapiro (2014) find similar fault with Goldman's interpretation of "embodiment" in "embodied cognition." I'll say more below about this and about whether Goldman bears some blame for his interpretation of "embodiment." I'll suggest that he does not.

3. Kiverstein's Own View

While Kiverstein rejects Goldman's interpretation of "embodiment" concerning the "neural reuse" hypothesis, he accepts Anderson's interpretation (2010, 2014, 2016). He accepts Anderson's because he believes Anderson's supports the strong interpretation of embodiment. That is, he thinks Anderson's neural reuse hypothesis supports the idea that cognition is partially constituted by representations in perceptual areas and/or motor areas of the brain. If true, this would support the strong interpretation of embodiment and would justify the excitement over this new idea in cognitive science. But is it true? Does Anderson's hypothesis and evidence support the strong view of embodiment? I think the answer is no.

First, it is clear that Kiverstein is aware of the distinction I'm making between mere causal support of cognition and what constitutes cognition. In footnote number 9 he says this: "Proponents of an amodal theory of symbolic thought can allow, for instance, that often when concepts are tokened in thought there is accompanying or associated visual or motor imagery. It doesn't follow that this accompanying imagery plays a constituting role in conceptual thought."

So it is clear that Kiverstein acknowledges the distinction I am pressing.

Kiverstein claims that Goldman's notion of embodiment is too narrow. He says:

> A science of embodied cognition does not only cover cognitive processes that make use of bodily formatted representations, it aims to explain how the entire spectrum of distinctively human capacities for higher-order cognition could have evolved through the reuse and repurposing of more basic sensorimotor capacities.
>
> (p. 89)

Second, while this is quite true, it still doesn't explain whether the reuse is causally relating sensorimotor processing to cognition or making these processes constitutive of cognition. So, it still doesn't show that Goldman's weaker view is wrong, nor show that the stronger view of embodiment is true.

Third, what Kiverstein says he likes about Anderson's view is that:

> Anderson's analysis of embodied cognition allows by contrast that many cognitive processes don't only take place in the brain, but in the sensorimotor dynamic coupling of the agent with its ecological setting. It is the body itself in its interactions with a richly resourceful environment that is involved in

> the performance of cognitive tasks, over and above any representations of the body that might be involved in the performance of such tasks.
>
> (pp. 89–90)

But involved how? Involved causally or constitutively? Does Anderson provide evidence of the stronger claim in his understanding of neural reuse, or only evidence for the weaker claim (consistent with Goldman's understanding)? The fact that more than representations are causally contributed by the body and environment to solving cognitive tasks does not show that the body and environment are involved constitutively. To show that takes significant argumentation.

Kiverstein gives an impressive list of quotes by several researchers detailing the contribution to cognition by parts of the body and environment, but not one of these authors makes the bold claim that these bodily and environmental contributions constitute cognitive processing. For all the quotes, their evidence can be interpreted as making causal contributions to cognition but not partially constituting cognition. Even the remark Kiverstein inserts by Dewey on environmental "mutualism" between organism and environment does not imply constitutional mutualism versus causal mutualism (p. 91). The same is true for remarks on "niche construction" (p. 92). Even the quote from an evolutionary perspective by Ingold about "the whole organic being (indissolubly mind and body) situated in a richly structured environment" doesn't address the weak versus strong view of embodiment (p. 93).

As for Glenberg and Kaschak (2002), even they note at the end of their paper that their findings are consistent with a weak view of embodiment. They note that the Action Compatibility Effects (ACEs) they discovered might be explained as post-cognitive effects—in which case the motor system activities would not be constitutive of cognition. Of course, they think the best interpretation of their data is that the motor system is involved in cognition, but they have nothing more in their theory or data that demonstrates this.[1]

Well what of Anderson himself? Surely he must make the bolder claim, since Kiverstein sides with his interpretation and against Goldman's. Well, I've read the quotes and descriptions Kiverstein offers again and again, and in them I can't find Anderson even addressing the issue of weak versus strong interpretation of embodiment, much less defending the strong interpretation of reuse. True: Anderson claims that the brain must be conceived of as an action-controller (p. 93). True: Anderson claims that the brain exploits environmental "affordances," and that the brain's use of these can be "repurposed," and suggests that selectional pressures can explain the changes (p. 93).

What about Anderson's discussion of "Transiently Assembled Local Neural Subsystems (TALoNS)"? Here is the conclusion Kiverstein draws from this part of his paper:

> The idea I want to tentatively explore next is that other higher-cognitive functions might be the result of putting together TALoNS widely distributed across the brain in ways that are tightly coordinated to the task at hand and the situation in which one is acting. Each TALoNS has an NRP [neuroscientifically

> relevant psychological] factor that relates to sensorimotor interaction and coordination with the affordances of the environment. However TALoNS can be combined in novel ways to support the kinds of higher-cognitive capacities that are distinctive of human cognition.
>
> (p. 97)

There is nothing here to interpret this research as supporting the strong (constitution) view over the weak (causal contribution) view. What is more, in this paper, Kiverstein does not address the point. Nor do the quotes from Anderson. Nor do the quotes from other researchers mentioned in conjunction with this concept.

At one point Kiverstein points to research on "pick," "kick," and "lick" by Pulvermüller (2005). This research demonstrates that when subjects in fMRI scanners hear these words, the areas of their brains that produce these movements fire. But does this show that to understand these words these brain areas must fire? No. And Pulvermüller is well aware of this and has offered several other conjectures to try to shore up the stronger interpretation (Puvermüller, 2012).[2] So, in effect, even here there is no argument for constitution given in Kiverstein's paper, nor in any of the data he cites in support of his stronger interpretation of the "reuse" hypothesis. For all of his criticism aimed at Goldman for not coming on strong for embodiment, Kiverstein himself is just as guilty of offering nothing stronger than support for the weak causal contribution view of embodiment.[3]

4. Goldman's Response to Shapiro and Kiverstein

Since to a large degree the complaint Kiverstein raises to Goldman is similar to that made by Shapiro (2014), I shall take Goldman's reply to Shapiro to apply equally to Kiverstein. As I pointed out above, among proponents of embodied cognition, there are those who support both weak and strong views of the role of the body in cognition (de Vega et al., 2012; Semin & Smith, 2008).[4] On the weak view, the body (including brain and motor areas, the so-called "modal" areas) provides causal support for cognition, but its inputs are not themselves constitutive of cognition. In the same way, the keyboard of a computer provides input to computation but key presses are not themselves computations. And similarly for output systems such as printer and video screens, these are causally connected to computation but signals sent to them are not themselves computations. So perceptual and motor system representations provide causal support to cognition but their activities do not themselves partially constitute cognition on the weak view. These are outside the "sandwich" of cognition, to use Hurley's metaphor.

On the strong view, activities in perceptual and motor areas of the brain (indeed even in the body outside the brain itself) might partially constitute cognitive processes, not just causally support them. On this view, the "sandwich" model is completely false. The body, perceptual areas, and motor areas contain cognitive activities. So on this view, so-called "modal" representations in these regions may themselves be cognitive representations. To think otherwise is thought to be a mistaken and outdated view of the nature of cognition and cognitive systems.

Does Goldman know these two separate interpretations exist? Yes; however, he chooses to avoid taking sides. He says:

> Unfortunately, the precise role of the body in this literature is mainly gestured at rather than clearly delineated. Moreover, it is difficult to pinpoint the empirical support for the theses that one can sink one's teeth into. These are among the principal reasons why I find them less satisfactory or persuasive than the styles of support for EC to be found in other branches of cognitive science, specifically cognitive psychology and neuroscience.
>
> (Goldman, 2014, p. 92)

It is specifically the matter of how to interpret the role of the body in embodied cognition (EC) that Goldman is referring to in the quote. He declines to take sides.

How can we be so sure that Goldman understands this distinction in interpretation? Sometimes he talks about the idea that for embodied cognition thought and language must be "grounded" in sensorimotor cognition. This is the kind of language used by Barsalou and Glenberg (in Semin & Smith, 2008). But of course, the issue is what does "grounding" mean? Does it mean causally supporting cognition or does it mean constituting cognition? It is ambiguous. Sometimes it is ambiguous even for folks like Barsalou and Glenberg, although they understand the two interpretations in play.

Goldman clearly understands the "sandwich view" of cognition which would make inputs from the body only causal contributions to cognition. Goldman says:

> The traditional idea of classical cognitivism is that pure, amodal cognition occupies a level of cognition entirely segregated from perception and motor execution … There are two entirely different ways of formulating and/or interpreting EC theses. On one interpretation, the body itself (and its various parts) plays a crucial role in cognition, a much more pervasive role than classical cognitivism recognizes.
>
> (Goldman, 2014, p. 93)

Indeed the "crucial" role is constitution. So he knows of this interpretation of embodied cognition. And he knows of the weaker view of EC, saying:

> On the second interpretation, it is representations of the body and its parts that are so pervasive and important to cognition. Theorists like Brooks (1999), Thelen and Smith (1994), and others contend that cognition is significantly mediated by the body's interaction with its environment, where this interaction does not take the form of the mind's *representation* of the body.
>
> (Goldman, 2014, p. 93)

Indeed, the weaker interpretation is that of mere causal support of cognitive processing provided by activities in the body outside the "sandwich." And for proof that Goldman knows of and uses the "constitution" language, he quotes from Mahon & Caramazza (2005) where they deny that current empirical data supports

the view that events in perception or motor regions "constitute" cognitive processing (Goldman, 2014, p. 98).

So what is really going on with Goldman? In my view, he knows about these two interpretations. He knows there is a dispute about which one is correct, but he is unwilling to take a stand. He told us (above) that there simply isn't enough data at this point to "sink one's teeth into" in order to decide.

Thus what is Goldman really doing, if he isn't entering into this dispute among EC theorists? He's doing something else. He is offering his own interpretation of what "EC" is supposed to mean. And then, given that meaning, he is using it for a different purpose (to help understand social cognition along the lines of Goldman's own "simulation theory" of mindreading).

Goldman offers this understanding of EC:

> If a cognition C uses an internal bodily format in the process of executing some cognitive task T, then even if task T is in no recognizable sense a bodily task (but rather a higher-level task of some kind), C still qualifies as an embodied cognition.
>
> (Goldman, 2014, p. 102)

This is what he means by "embodied." He simply is playing a different game than the one Shapiro and Kiverstein take him or want him to be playing. What is more, in reply to Shapiro, he acknowledges that some will be disappointed:

> I am fully aware that my proposal will leave many self-styled EC theorists unmoved and unimpressed. Their vision of the field is too remote from the one I wish to plow. So, I never had any illusion of being able to forge a union with them.
>
> (Goldman, 2014, p. 103)

Goldman continues:

> Shapiro's principal complaint, perhaps, is that my way of framing EC, with its adherence to brain-centrism, threatens to divest the embodiment movement of its most exciting and distinctive departure from orthodox cognitive science. Perhaps; but excitement isn't everything, especially in science.
>
> (Goldman, 2014, p. 104)

But is it wrong for Goldman to defend only a weak notion of embodiment? I cannot see why it would be wrong. As I say, many theorists who discuss embodied cognition never even mention the stronger reading, much less defend it (de Vega et al., 2012, Semin & Smith, 2008).

Now I think Shapiro makes the point that there must be a difference between what counts as embodied cognition and what does not, and Shapiro thinks this is a shortcoming of Goldman's account. But that would be to force everyone who claims to embrace the notion of "embodied cognition" to accept and defend the strong view. Why should one demand so much from the embodied view? The

role of the body could be significantly enhanced in the view of many cognitive scientists without them all accepting the strong interpretation of "embodied."

Perhaps it would be bad if Goldman didn't realize that *de facto* his view is the weak view of EC and that it is equivalent to the traditional view of cognition—the "sandwich" view in fact. It might be bad if he didn't realize that he is siding with classical cognition. It might be believed to be bad that he has turned his back on the main thing that makes EC new and exciting—viz., that it (the strong view of EC) is not just the classical view. But in the first place, Goldman does realize what he is doing and, in the second place, he admits and emphasizes that what he is doing will be found unexciting by many. So, I fail to see what he does that is wrong (not exciting, perhaps, but not wrong).

5. Conclusion

Kiverstein complains that Goldman's interpretation of "embodiment" is a mistake. He maintains that it is a mistake because it is too weak and is consistent with the traditional view of cognition, the "sandwich view" in fact (though Kiverstein does not use Hurley's words). However, many theorists who discuss embodied cognition do not properly sort out the differences between the weak and strong views. Many give scientific data that supports only the weak view or at least does not differentiate the weak view from the strong view.

Goldman, to his credit, acknowledges the differences between the strong and weak interpretations of "embodied." He offers a theory of neural reuse that supports only the weak view of embodiment. He acknowledges this very fact. Both Kiverstein and Shapiro are unsatisfied with this weak understanding. Kiverstein is unsatisfied because he wants to defend the stronger view and finds it more to his liking. Shapiro is unsatisfied because he fears the weaker view threatens not to differentiate embodied from non-embodied views. And Shapiro is right—but Goldman is openly confessing to this very fact.

What is more, with the distinction between strong and weak notions of embodiment out in the open, Kiverstein's complaints against Goldman ring hollow. Neither Kiverstein nor any of the researchers he cites for support mount an argument in defense of the strong interpretation of embodiment over the weak. It is clear that Kiverstein wants the reader to take the support offered as evidence of the strong reading, but the evidence is equivocal. It could be taken either way. So if Kiverstein wants to defend the strong (constitutive) role of the body in cognition, then he will have to supply much stronger support in the future. I will be eager to see what he has to offer.[5]

Notes

1 have discussed this in person with Glenberg. He understands the weak versus strong readings of "embodiment" but he has not to my knowledge addressed the matter in print.

2 believe these other considerations fail, but won't discuss them here. The point here is that neither Kiverstein nor any one of the researchers he cites here offers support for the strong interpretation of embodiment.

3 See also Adams (2010).
4 The authors in these texts rarely clearly differentiate the strong from the weak readings and hence it is often very difficult to tell which view they are defending.
5 Thank you very much to Kiverstein and the editors for inviting me to contribute to this volume. I would also like to thank Charlotte Shreve for very helpful conversation and advice.

References

Adams, F. 2010. Embodied cognition. *Phenomenology and the Cognitive Sciences*, 9(4), 619–28.
Adams, F., & Aizawa, K. 2008. *The Bounds of Cognition*. Oxford: Wiley/Blackwell.
Anderson, M. L. 2010. Neural reuse: A fundamental organisational principle of the brain. *Behavioural and Brain Sciences,* 33(4), 245–66.
Anderson, M. L. 2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.
Anderson, M. L. 2016. Precis of *After Phrenology*: Neural Reuse and the Interactive Brain. *Behavioural and Brain Sciences*, 39, e120. doi: https://doi.org/10.1017/S0140525X15000631
Clark, A., & Chalmers, D. 1998. The Extended Mind. *Analysis*, 58, 10–23.
de Vega, M., Glenberg, A., & Graesser, A. 2012. *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford: Oxford University Press.
Glenberg, A. M., & Kaschak, M. P. 2002. Grounding language in action. *Psychonomic Bulletin and Review*, 9, 558–65.
Goldman, A. 2014. The bodily formats approach to embodied cognition. In U. Kriegel (Ed.), *Current Controversies in the Philosophy of Mind*. London: Routledge.
Hurley, S. 2001. Perception and action: Alternative views. *Synthese*, 129, 3–40.
Mahon, B. Z., & Caramazza, A. 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology—Paris*, 102(1), 59–76. doi: https://doi.org/10.1016/j.jphysparis.2008.03.004
Pulvermüller, F. 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6, 576–82.
Pulvermüller, F. 2012. Grounding language in the brain. In M. de Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford: Oxford University Press, pp. 85–116.
Semin, G. R., & Smith, E. R. 2008. *Embodied Grounding: Social, Cognitive, Affective and Neuroscientific Approaches.* Cambridge: Cambridge University Press.
Shapiro, L. 2014. When Is Cognition Embodied? In U. Kriegel (Ed.), *Current Controversies in Philosophy of Mind*. New York: Routledge, pp. 73–90.

Further Readings for Part III

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4), 245–66. https://doi.org/10.1017/S0140525X10000853

A widely cited article argues that neural circuits can be exapted for new uses while retaining their original uses, as theories of embodied cognition claim. Includes a number of insightful peer commentaries and replies from the author of the target article.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, 3(1), 71–88. https://doi.org/10.1007/s13164-012-0089-0

Argues for a view on which representations using a particular bodily code play an important role in cognitive functions beyond representing actions and states of the cognizer's own body.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology—Paris*, 102(1), 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004

A widely cited critique of arguments for embodied views of conceptual content.

Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic Bulletin & Review*, 23(4), 941–58. https://doi.org/10.3758/s13423-016-1045-2

Introduces a special issue on conceptual representation that contains a number of articles on embodied cognition.

Newen, A., Bruin, L. D., & Gallagher, S. (2018). *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Contains many contributions on embodied cognition. See especially the introduction and chapters in Part I, Part IV, and Part V.

Niedenthal, C., Wood, A., & Rychlowska, M. (2014). Embodied emotion concepts. In L. Shapiro (Ed.), *The Routledge Handbook of Embodied Cognition*, 240–9. New York: Routledge.

A concise survey of the evidence that representations of emotions are embodied representations.

Scorolli, C. (2014). Embodiment and language. In L. Shapiro (Ed.), *The Routledge Handbook of Embodied Cognition*, 127–38. New York: Routledge.

A concise survey of the evidence that embodied representations play an important role in language processing.

Shapiro, L. (2010). *Embodied Cognition*. London: Routledge.

An introduction to classic debates in the philosophy of cognitive science about the relationship between embodied cognition and traditional approaches to cognitive science.

Study Questions for Part III

1) What distinguishes Goldman's less radical theory of embodied cognition from Kiverstein's more radical theory of embodied cognition?
2) According to Kiverstein, why does the data reviewed by Anderson support his more radical theory of embodied cognition over Goldman's less radical theory of embodied cognition?
3) According to Adams, why does the evidence adduced by Kiverstein fail to support his theory over Goldman's less radical theory?
4) What is the difference between "weak" forms of embodiment and "strong" forms of embodiment?
5) Could a theory of embodied cognition be strong while being consistent with the traditional computational theory of mind? If so, why? If not, why not?

Part IV

How Should Neuroscience Inform the Study of Cognition?



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7 Is Cognitive Neuroscience an Oxymoron?

Fiery Cushman

1. Introduction

Cognitive neuroscience research requires a great deal of time and money, and it attracts a great deal of attention. Consequently, there is a subversive temptation to declare it useless. This general grumble comprises many related gripes: Cognitive neuroscience is poorly executed; its costs do not justify its benefits; for a long time we all got along just fine without it, and I still do; it's just a fad; people are just wowed by pictures of brains, etc.

Still, cognition is the study of mental computations that are performed by a neural machine. It is remarkable to suppose that studying the machine itself would not usefully inform cognitive research. So which is it? Does neuroscience illuminate cognitive research, or is it the moon that never rose? Is cognitive neuroscience an oxymoron?

This chapter dodges numerous gripes. It does not address whether most cognitive neuroscience is well-executed; whether its benefits typically outweigh its costs; whether neuroscience could ever be necessary to develop a psychological theory; or whether its celebrity outstrips its merit. But it does conclude that neuroscience can play an important role in cognitive research. The argument has three parts: a general account of the relationship between neuroscience and cognition, a specific case study of the contribution of neuroscience to the study of visual cognition, and a quantitative analysis of citation records in the journal *Cognition*.

1.1 What Is Cognitive Research?

Before asking whether cognitive neuroscience contributes to cognitive research, it helps to have a clear image in mind of just what "cognitive research" is. Cognitive theories attempt to describe the functional organization of the mind (Gallistel, 1990). The guiding principle of the cognitive revolution is that we can understand how the mind processes information. Specifically, we can characterize mental activity as a set of representations and computations over those representations. This depends crucially upon an understanding of the function of the mental processes under study (Marr, 1982).

Some classic examples of psychological research questions formulated in this framework concern visual object recognition, language production, and theory of

mind. In the case of visual object recognition, for instance, it is assumed that the mind begins with a pixel-like representation of the wavelengths and luminosities of light across the retina. It ends with the capacity to categorize objects in the visual scene—e.g., categorizing one as a CAT and another as a HAT. This occurs as successive computations are performed on the early representation, transforming it over a series of representations.

This commitment to characterizing the internal representations that mediate between percept and action is precisely why behaviorists found the cognitive revolution so revolting. Behaviorism is committed, of course, to merely characterizing the relationship between a stimulus and a response. This is a kind of psychological theory. Indeed, it aptly describes the kinds of theories of human behavior applied in some other disciplines. For instance, several key principles of microeconomics do not attempt to mirror the representations or computations involved in choice, except insofar as they appropriately characterize the relationship between input and output of a human chooser over time. Such theories achieve descriptive adequacy, in the sense that they accurately describe and predict human behavior (Chomsky, 1965). Cognitive theories, in contrast, achieve *explanatory* adequacy in addition. That is, they aim to characterize the set of internal representations and computations that provide an explanation for the reliability of descriptive relations between input and output. Thus, if neuroscience plays an important role in cognitive research, it must do so by constraining theories of representation and computation.

1.2 What Is Neuroscience?

Neuroscience is the study of the nervous system, including its physical and biological properties. Like all cognitive theories, many neuroscientific theories are motivated by the ultimate desire to explain the functional organization of the mind. Yet theories of neuroscience can be distinguished from cognitive theories in part because they are stated over physical and biological terms (or explicit abstractions from them) rather than exclusively representational and computational terms.

Neuroscience comprises not only a set of theories, but also a set of methods. A neuroscience method involves either manipulating or measuring neural processes directly. In other words, either it involves an intervention upon neural processes in a manner that is not mediated by ordinary perception (e.g., neural stimulation by electrical or magnetic forces, lesion by natural or artificial means, experimental control of gene expression, etc.), or else it involves measuring neural processes in a manner that is not mediated by ordinary behavior (e.g., EEG, MRI, single-unit recording, etc.). In contrast, in "behavioral" methods, the experimenter's access to mental processes is mediated by ordinary perceptual and behavioral processes.

It is not necessary that neuroscience methods are used exclusively to inform neuroscience theories, or that behavioral methods are used exclusively to inform cognitive theories. This point is crucial: If we defined the methods in terms of the theory they seek to advance, this would beg the question of whether neuroscience methods (e.g., cognitive neuroscience) can inform cognitive theory. The answer would be guaranteed to be "no" for a trivial reason: Any study using apparent

"neuroscience" methods to answer cognitive questions would simply have been "cognitive" research all along, by definition.

1.3 What Is "Important"?

What does it mean to say that neuroscience has played, or could play, an "important" role in the development of cognitive theories? This question is most interesting if it is reframed pragmatically: If we are interested in developing an accurate model of cognition, is it a useful strategy to seek evidence from neuroscience methods?

Thus, neuroscience could be a useful source of evidence without being *sufficient* for the development of cognitive theories. In other words, having a complete theory of cognition may require us to know about more than just the brain. A complete cognitive theory of visual processing, for instance, might depend upon information about typical scenes in natural contexts (i.e., the things we see), as well as the functional role of vision in guiding human interactions in the world. Whether cognitive theories must be stated over terms external to the nervous system is a matter of significant philosophical controversy. But, even if we adopt the position that a theory of mental representation and computation must be stated partially over terms external to the nervous system, neuroscience could still play a useful role in developing those theories.

Similarly, neuroscience could be a useful source of evidence in developing cognitive theories without being a *definitive* source of evidence. (Indeed, it's hard to say whether any form of scientific evidence is ever definitive.) Presumably neuroscience will constrain cognitive theories via inference to the best explanation. That is, certain cognitive theories will provide a better explanation for the neuroscience evidence, others will provide a worse explanation, and we can use this to inform our assessment of the quality of the cognitive theories. In this method of reasoning, the neuroscience evidence can play an important role in adjusting our assessment of the probability of cognitive theories being correct, even without definitively endorsing or defeating any particular theory.

Neuroscience could also be a useful source of evidence without being *necessary* for the development of cognitive theories. This point is obvious enough. If I want to know what color my socks are, looking at them would be a useful source of evidence, but it is certainly not a necessary source of evidence. I could buy a camera, hire a film crew to shoot a movie of my socks, show it to my wife and ask her what color she sees. But even if it isn't necessary for me to look at my own socks to determine their color, it could certainly be useful.

Finally, it might be the case that neuroscience is a possible source of evidence without being a practical one. For instance, the film-crew approach is a possible method of obtaining evidence about the color of my socks, but it isn't a very practical one. Although practicality is important, I won't address it much further.

2. *A Priori* Analysis

The brain is the principal physical substrate of human cognition. Doesn't this guarantee that theories of brain function will be relevant to theories of cognition?

In fact it does not, as a simple analogy illustrates. Consider a scholar who wishes to develop a theory of the influence of *Uncle Tom's Cabin* on American attitudes towards slavery. This novel was distributed in printed books, and so the influence of its content upon readers was fully mediated by physical objects. Would it be useful, then, for the scholar to investigate properties of those objects, such as their material composition (paper and ink) or method of manufacture? Probably not: This approach would fail because general information about the material composition of the book or its method of manufacture are not sufficiently related to the information contained in the book, which is what shaped American attitudes towards slavery.

The same is true for many properties of neural systems when applied to cognitive theory. For instance, there are detailed theories of the cellular mechanisms that enable neuronal firing; however, these are unlikely to be sufficiently related to mental representation and computation to play an important role in cognitive research. Even if every cognitive representation is encoded by the firing of neurons, this does not mean that a theory of the mechanics of neural firing at the cellular level informs a theory of cognitive representation.

Consider a slightly different question, however: If a scholar wishes to understand the influence of *Uncle Tom's Cabin* on American attitudes towards slavery, would it be useful for her to read the book—i.e., to interact directly with it as a physical object? Presumably it would, simply because this would provide relevant information about its contents, which would in turn inform a theory of its influence. Reading the book might not be necessary. The scholar could rely on indirect sources of evidence about its content, and many of the inferences she would wish to draw about its influence wouldn't depend on knowing its precise contents anyway. Nor, of course, would reading the book be sufficient to answer all of the scholar's questions. But if the scholar had never read the book before, and if it sat on her shelf, surely she would be crazy not to pick it up.

This analogy highlights the utility of distinguishing between neuroscience *theory* and neuroscience *methods*. A general theory of "books" as a physical object might be stated over terms irrelevant to the informational content of any given book—e.g., principles of chemical bonding, methods of manufacture, etc. Such theories would typically be useless to a scholar interested in their informational content. Yet, the "method" of interacting physically with particular books (i.e., reading them) is obviously a very useful way of understanding their informational content. Similarly, it might be the case that theories of neuroscience typically generalize over properties of neural systems that are not particularly informative about mental organization at the level of representation and computation. Yet, interacting directly with the physical substrate of a neural system might be a very useful method of learning about the representations and computations instantiated in one.

Of course, reading the brain is not as easy as reading a book. If your uncle Tim is acting oddly and you want to understand why, it probably makes more sense to use behavioral methods (i.e., observing the relationship between perceptual inputs and behavioral outputs) than to crack open his head and take a close look at his brain. In order for the "direct" investigation of the neural system to usefully

inform psychology, we must have theories capable of relating physical data to cognitive theories, and instruments capable of collecting that data. Given such theories and instruments, however, it is hard to see how neuroscience methods could fail to be informative for cognitive research.

3. A Case Study

A classic cognitive research problem is to understand the information processing that enables vision. Somehow, we transform a dynamic two-dimensional projection of light onto the retina in each eye into a three-dimensional representation of space, a categorization of objects within that space, and a representation of the motion of those objects over time. In cognitive terms, the goal is to understand the nature of the information represented on the retina, and then to specify the series of computations that transform this information across a series of representations, resulting finally in representations of three-dimensional space, object identity, motion, etc.

Such a cognitive theory can be neutral with respect to neural mechanism. For instance, consider the hypothesis that a key step during object recognition is the discovery of object boundaries by computing local luminosity gradients. By itself, such a hypothesis is not committed to any particular neural mechanism; indeed, it applies as naturally to visual processing on a laptop computer as it does to visual processing in a brain. Yet, over a period of about 20 years, neuroscientists actually did drive pioneering advances in our understanding of the earliest stages of visual processing. Now several decades later, we are in a good position to appreciate the historical significance of this work, which is an ideal case study of the potential for neuroscience to advance cognitive research.

Feed-forward visual processing begins when light hits a photoreceptor in the retina. These photoreceptors relay electrical signals through another layer of cells to eventually activate retinal ganglion cells, which transmit information from the eye to the brain. In the years surrounding World War II, a number of researchers performed single-unit recordings of the evoked response of retinal ganglion cells during the presentation of brief visual stimuli to the eye (H. K. Hartline, 1938; H. Hartline, 1940; Barlow, 1953; Kuffler, 1953). This line of investigation led to a pivotal characterization of a type of receptive field common to many retinal ganglion cells: "center surround" (Figure 7.1).

An immediately subsequent stage of processing was characterized in a classic series of experiments by Hubel and Wiesel (1962). This stage takes representations with a center-surround format as input. By computing the spatial relations among cells of this type, it derives a representation of local linear contrast boundaries and their orientation (Figure 7.2). Here, again, the experiments that established the representational content of early visual processing were single-unit recordings, in this case taken from the lateral geniculate nucleus and cortical area V1.

These experiments had a transformational impact on the field because they provided a case study of the mapping between neural function and cognitive organization (Gordon, 2004; Wurtz, 2009). Experimental evidence characterized the representational content of three successive stages of visual processing,

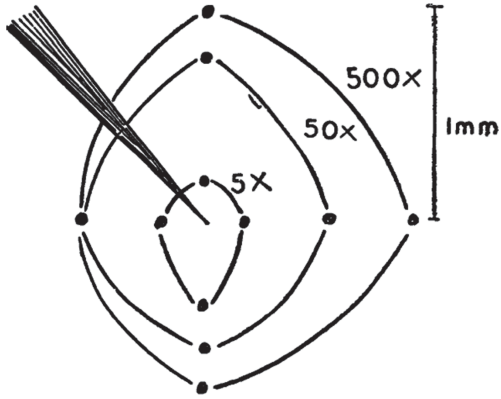


Figure 7.1 A reproduction of an illustration from Kuffler (1953) showing the receptive field of a retinal ganglion cell. In the diagrammed receptive field, the innermost portion responds preferentially to the onset of light, while the outermost portion responds preferentially to the offset of light.

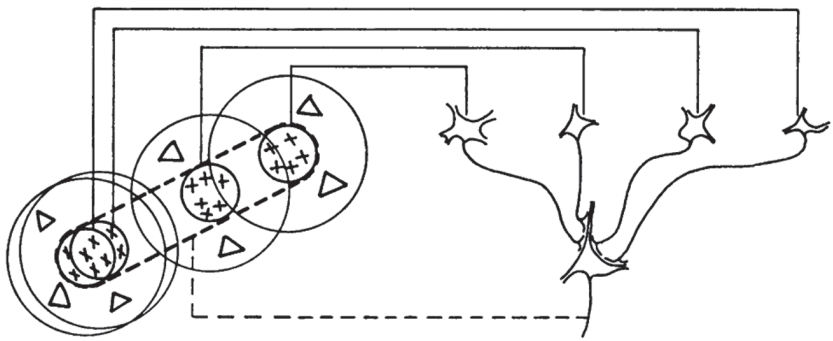


Figure 7.2 A reproduction of an illustration from Hubel and Wiesel (1962) showing how the activation levels of several center-surround cells could be summed in order to produce the receptive field of a complex cell.

specifically in terms of the activation of photoreceptors and the receptive fields of individual neurons across two populations. And a biologically plausible model showed how connections between cells and their individual response properties could instantiate the computation of the later representations from earlier ones. Perhaps most importantly, it was immediately obvious that this element of visual processing constituted useful progress towards the functional purpose of object recognition. Specifically, a retinal map of luminosity was being transformed into elements useful for the detection of object edges in natural scenes (which often involve local high-contrast boundaries). Later cognitive

theories posited that this process of edge detection plays a crucial role in object identification and the construction of a three-dimensional representation of space (Marr, 1982).

Notably, in this case the neural evidence preceded the development of a cognitive theory. Researchers had not specified a hypothesis of visual processing that included representations of center-surround receptive fields. To the contrary, similar experiments performed on the frog's retinal ganglion identified representational contents that long confounded any attempt to relate them to the functional demands of vision (a problem famously addressed in Lettvin, Maturana, McCulloch, & Pitts, 1959).

During this formative period of the study of visual processing, then, neuroscience methods drove breakthroughs in cognitive theory. Hubel and Wiesel's experiments were examples of neuroscience methods because their dependent measure was a physiological signal produced by the brain, rather than behavior. Yet, these same experiments were designed to build a cognitive theory of vision. Their basic strategy, which was to characterize the response profile of neurons (i.e., their receptive fields), is reflected across a vast subsequent literature, and encompasses not just single-unit recording but also EEG, PET, MRI, etc. (DiCarlo, Zoccolan, & Rust, 2012).

Indeed, an immediate impact of Hubel and Wiesel's research was to prompt new directions in behavioral and computational research on early visual processing. The receptive fields of some LGN and V1 neurons characterized by Hubel and Wiesel resemble a local sinusoidal contrast gradient. These are distributed at variable orientations and over variable spatial frequencies (i.e., variable "widths" of the sine wave). It was soon noted that a population of neurons with receptive fields of this type would be capable of representing a visual scene by approximating a Fourier decomposition of the spatial frequency of luminosities. Eventually this line of inquiry established a correspondence between the receptive fields of complex cells and Gabor wavelets, a mathematical object that enables an efficient compression of the information contained in natural scenes. In fact, the basic strategy of decomposing natural scenes into summaries of local spatial frequency is shared with common image compression algorithms for computers, such as the JPEG format.

As this computational interpretation of the neuroscience was developed, a number of clever behavioral paradigms were devised to demonstrate its validity (e.g. Campbell & Robson, 1968). Thus, in all likelihood, the same cognitive theory *could* have been developed without constraint or inspiration from neuroscientific evidence. Even if a suitably precise cognitive theory of visual processing could be developed without evidence from neuroscience, however, it is easy to grasp how pivotal the role of neuroscience evidence was in this case. A characterization of the receptive fields of individual neurons across multiples stages of representation enabled Hubel and Wiesel to posit an account of the representations and computations contained in early visual processing that had not really been entertained in any serious way by prior theorists. In other words, they made a big step, relatively fast, and there is little indication that the step would have soon occurred without reliance on neuroscience.

Above, I introduced the conceptual distinction between neuroscience theory and neuroscience methods. The key contribution of Hubel and Wiesel's research to a cognitive theory of visual processing appears to have been their use of neuroscience methods, not their development of a general theory of neuroscience. In other words, the information they obtained about neural responses did not influence cognitive theory by way of some general theory of the nervous system—i.e., a theory that generalized over physical rather than informational properties of the brain. Hubel and Wiesel instead moved directly from data about the neural system, collected by neuroscience methods, to a theory of information processing. Although their theory of information processing could also be construed as a "neuroscience theory"—after all, it described the implementation of an algorithm in terms of the firing of and connections between neurons—it was no less a theory of cognition. By marrying an algorithmic specification of cognitive processing to a biologically plausible neural implementation (Marr, 1982) it was, without irony or contradiction, an instance of true "cognitive neuroscience."

4. Quantitative Analysis

If neuroscience is important to the development of cognitive theories, then cognitive research that is directly informed by neuroscience will tend to be of a superior quality. How can we test this hypothesis by quantitative analysis? A common proxy for the quality of scientific research is its citation index. And cognitive research that is "influenced" by neuroscience will, of course, tend to cite that very neuroscience. Thus, we might ask whether cognitive research that *cites neuroscience* is subsequently *more cited*. This evidence would be consistent with the hypothesis that cognitive research that is influenced by neuroscience is of a higher quality (or at any rate more influential), although it is also consistent with some alternative hypotheses.

Of course, it is crucial that this hypothesis is tested on papers that are unambiguously devoted to the development of cognitive theory. It is possible, for instance, that neuroscience research is more cited than cognitive research, but this is not the question. Our goal is not to establish whether neuroscience is influential in general. Rather, the goal is to assess whether neuroscience improves *cognitive research*. Thus, we want to consider only articles that contain cognitive research, asking whether those that cite neuroscience (i.e., are influenced by it) are subsequently more cited (i.e., are of a higher quality, or more influential).

We assessed this question in a sample of articles drawn from the journal *Cognition*. This journal publishes articles almost exclusively on cognition. It publishes almost no research employing neuroscience methods (e.g., fMRI), and the few articles that do employ neuroscience methods are designed to address theories of cognition. Indeed, among the sample of articles that we targeted, a search on the Thomson Reuters Web of Science did not reveal any that included "fMRI" in its title or abstract. The audience of *Cognition* is primarily composed of psychologists, and it is among the more influential and widely read journals specific to the field of cognitive psychology.

This analysis focuses on articles published during the calendar years 2008, 2009, and 2010. This affords a large sample of published articles (N = 540) for which we could obtain adequate citation information from the Web of Science. We chose this range of dates because it seemed late enough for the published articles to have potentially benefited from the explosive growth of cognitive neuroscience in the preceding decade, but early enough to have amassed a meaningful number of citations representative of their quality.

For each article, we recorded the number of Web of Science indexed citations (hereafter "citations") through August 2015 (the date at which this manuscript was initially prepared); this is our measure of the article's quality and influence. We used an automated text extraction algorithm to identify the references cited in each article PDF (hereafter, "references"). This algorithm identified 21,159 references, including many to the same sources. We then categorized references targeting "neuroscience" (versus not) by assessing whether each referenced article was published in any of 62 neuroscience journals, or whether its title contained word stems associated with neuroscience (e.g., "Neur", "Brain", etc.).

The sample of articles we analyzed were cited an average of 31 times each. Out of 540 total articles, 311 (58%) referenced at least one neuroscience article, while 229 (42%) referenced none. Articles that contained at least one neuroscience reference received an average of 34 citations, while those that contained no neuroscience references received an average of 26 citations, which is a statistically significant difference $t(538) = -2.88$, $p < .005$. In other words, articles that referenced neuroscience tended to receive about eight more citations on average than articles that did not, an increase of 31% in the frequency of citation. A comparison of the most cited to least cited articles helps to illustrate this trend. Among the 20 most cited articles (mean = 148 citations), 17 contained at least one neuroscience reference. In contrast, among the 20 least cited articles (mean = 3 citations), only 9 contained at least one neuroscience reference.

This analysis can be improved in a few simple ways. First, citations are not distributed normally across this sample of articles; rather, the distribution is highly right skewed, violating an assumption of the parametric statistical tests we employed to analyze our data. A log transformation of citation counts produces an adequately normally distributed dependent variable. Second, as might be expected, articles published earlier (e.g., 2008) tended to receive more citations than articles published later (e.g., 2010), while the likelihood of neuroscience references also increased over this range in our sample. This pattern of correlations may obscure the predicted effect of neuroscience references on subsequent citations, and so it is important to statistically control for the effect of time. Third, articles that contained more references overall were more likely to contain at least one reference to neuroscience, and also to receive more citations (e.g., perhaps because of variable rates of citation across subdisciplines, reciprocity among authors, a correlation between an exhaustive literature review and article quality, etc.). This pattern of correlations may artificially inflate the predicted effect; again, it is important to statistically control for it.

Thus, we performed a regression that predicted the natural logarithm of the number of citations received by each article in our sample. We included three

predictors: a dichotomous variable that coded whether each article contained any references to neuroscience, a variable coding the year of publication, and the total number of references contained in each article. Again, we found a statistically significant effect indicating that articles that referenced neuroscience were more likely to receive subsequent citations, $p < .005$, after controlling for the other two predictors.

Although we find clear evidence in this model that referencing *any* neuroscience is correlated with a higher citation index, we do not find evidence that referencing *more* neuroscience helps. Specifically, after controlling for the presence versus absence of any neuroscience reference, there is no further statistical relationship with the specific number of neuroscience references ($p > .50$) or the proportion of neuroscience references out of all references ($p > .50$). On the one hand, this provides some evidence against the hypothesis that articles referencing neuroscience are cited more only when they actually *contribute* to the neuroscience literature (i.e., report neuroscientific results, address neuroscientific theories, etc.). Were this the case, presumably an article that cites many neuroscience findings (and thus more likely contributes to the neuroscience literature) would receive more citations than an article that cites only one or two neuroscience findings (and thus likely does not contribute to the neuroscience literature). On the other hand, if neuroscience improves cognitive research, it is surprising that we do not find evidence that more neuroscience is associated with greater improvement.

Although the evidence described above is consistent with the possibility that neuroscience improves cognitive research, there are several important alternative explanations for the observed effect. One possibility is that it may be explained by variability across subdisciplines of cognitive research. For instance, research into cognitive development may infrequently refer to neuroscience (e.g., because little neuroscience is performed on infants and children) and also receive fewer citations (e.g., because the pace of research is relatively slow), compared with other subdisciplines. This would produce a spurious confound between neuroscience references and subsequent citations.

We conducted a quick and rough test of this account, testing for the presence of the basic "neuroscience advantage" already identified, but now within four separate categories of research. We identified articles belonging to each research category by searching for keywords or word stems in the titles of articles contained in our sample. For instance, twelve articles contained the word "moral" in the title, and thus presumably focused on moral cognition. Eight of these articles referenced neuroscience, and these received an average of 92 citations each; four did not reference neuroscience, and these received an average of 13 citations each. Among 25 articles with the word "face" in the title, those referencing neuroscience received more citations (mean = 29) than those that did not (25). Among 60 articles with the word stems "ling" or "lang" in the title, those referencing neuroscience again received more citations on average (42 vs. 22). The same trend emerged among 58 articles with the word stem "child" or "infant" in the title (35 vs. 32). These are the only categories we analyzed; clearly, further and more systematic approaches could be taken to address this question. Yet, while the magnitude of the "neuroscience

advantage" is variable across categories, in each case there is a trend in the same direction. This provides some preliminary evidence that the overall effect is not spuriously driven by features specific to different subdisciplines.

Another explanation for the observed pattern of results is that references to neuroscience impart a false aura of quality. In other words, people may cite articles that reference neuroscience because they have the impression that the research or theory is more cutting-edge, valid, interdisciplinary, plausible, etc., but in fact those articles may not differ in any of these respects. This explanation is difficult to rule out because of our use of citation as a proxy for quality. It may be amplified by our reliance on a sample of recently published articles; perhaps the "early returns" of citations are driven less by actual research quality than by heuristic approximations of it. The foundations of this argument are more than hypothetical: Perhaps ironically, the sixth-most-cited article in our sample is titled "Seeing is believing: The effect of brain images on judgments of scientific reasoning" (McCabe & Castel, 2008). It shows that college students are more likely to accept a scientific argument as valid if it is accompanied by an uninformative fMRI activation map. On the other hand, other research provides some evidence that scientific experts are immune to this effect (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008).

A final alternative explanation for this effect is that articles referencing neuroscience are more likely to be cited by neuroscience, but not by other branches of cognitive science. Why might this be? An unlikely possibility is that articles in the journal *Cognition* that reference neuroscience are being cited by neuroscientists not because of their contributions to cognitive theory, but rather because of their contributions to non-cognitive theories of neuroscience—e.g., a model of network dynamics, or neurotransmitter receptor types, etc. This possibility is unlikely because *Cognition* just doesn't publish articles with this kind of content. (Recall that not a single article in our sample contains "fMRI" in the title or abstract.)

A second possibility is that cognitive neuroscientists are more likely to cite articles (or authors) who cite them, perhaps via an implicit or explicit reciprocity, or because they find it easier to relate such work to their own interests. This hypothesis is more likely, and also harder to rule out based on our data. Because the citation advantage that we observe is not eliminated when controlling for the total number of citations, we can at least provide some evidence against a pure effect of peer-to-peer reciprocity in citation (assuming that neuroscientists are no more prone to such reciprocity than non-neuroscientists), but other versions of the hypothesis are not addressed by this analysis.

There is a third version of this alternative that is perhaps the most likely, but is friendly to the overall claim that cognitive research is improved by engagement with neuroscience. Possibly, cognitive research that engages seriously and directly with neuroscience is also more useful to neuroscientists who are working on cognitive questions—i.e., cognitive neuroscientists. For instance, research that uses behavioral methods to investigate early visual representations would likely benefit from research on this topic using neuroscience methods; in turn, it is especially likely to inform future research using neuroscience methods. In this case, our effect might be driven by citations from neuroscientists, and yet it would still be

diagnostic of the underlying quality and broad applicability of the research for the development of *cognitive* theories.

Based on the set of citation records available to us we could not develop a reliable method to directly assess whether our effect was due entirely to enhanced rates of citation by neuroscience for those articles in our sample that referenced neuroscience. The most relevant evidence we found targets a different but related question: Are references to neuroscience related to increased citation from articles that are *not* neuroscience? In order to assess this, we analyzed the number of citations for each article in our sample, focusing exclusively on subsequent articles also published in *Cognition*. In other words, we asked: "For three years' worth of articles published in *Cognition*, if they referenced neuroscience, were they then more likely to be cited by subsequent articles also published in *Cognition*?" In this case we found no significant difference in citation rates; indeed, if anything the citation rates trended in the opposite direction (i.e., more citations for articles that did not refer to neuroscience). This analysis, however, is compromised by a marked reduction in statistical power.

In summary, our analysis provides strong evidence that references to neuroscience articles are associated with higher citation rates for cognitive research in *Cognition*. Yet, this data is compatible with a number of rival hypotheses and does not provide exclusive support for the conclusion that neuroscience has an important influence on cognitive research.

5. Conclusion

Cognitive neuroscience aims to characterize the interface between computation and biology, and it is not an oxymoron. *A priori*, there is good reason to believe that cognitive neuroscience will advance cognitive theory. Although it is not clear whether "theories" of neuroscience will often play this role, there is every reason to believe that neuroscience methods will be pivotal. This depends, of course, on the existence of theories that map between physical properties of the brain and computational properties of the mind. A case study of an early advance in cognitive neuroscience, the mapping of receptive fields in the retinal ganglion and early visual cortex, shows that this approach can succeed in practice. This case study underscores the utility of neuroscience methods for describing representations and computations that are many stages of processing removed from behavior and, thus, difficult to characterize by behavioral methods alone. Of course, this case study may be anomalous. Yet, in an analysis of a sample of articles developing cognitive theories, we find that references to past neuroscience research are associated with a greater impact as measured by citation index. This is at best an indirect proxy for the underlying "quality" of cognitive research, however, and the evidence presented here is consistent with several alternative hypotheses.

Each of the methods used above—*a priori* analysis, case study, and quantitative analysis—has its own drawbacks. Together, however, their complementary strengths provide a unified account of the important role that evidence from neuroscience can play in advancing cognitive research.

Acknowledgments

Thanks to Matthew Cashman, who played a pivotal role in compiling the database of citation records used in this article, and to members of the Moral Psychology Research Lab for valuable feedback.

References

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *Journal of Physiology*, 119(1), 69–88.

Campbell, F. W., & Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, 197(3), 551–66.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–34.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Gordon, I. E. (2004). *Theories of Visual Perception*. Hove, UK: Psychology Press.

Hartline, H. (1940). The effects of spatial summation in the retina on the excitation of the fibers of the optic nerve. *American Journal of Physiology*, 130(4), 700–11.

Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology*, 121(2), 400–15.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1), 106.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1), 37–68.

Lettvin, J., Maturana, H., McCulloch, W., & Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11), 1940–51.

Marr, D. (1982). *Vision*. New York: Freeman.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, 107(1), 343–52.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20(3), 470–7.

Wurtz, R. H. (2009). Recounting the impact of Hubel and Wiesel. *Journal of Physiology*, 587(12), 2817–23.

8 On the Primacy of Behavioral Research for Understanding the Brain

Yael Niv

The advent of increasingly more precise methods for measuring and perturbing neurons in the brain has been accompanied by a demotion of animal and human behavioral research—the mainstays of psychology—to being "only" behavioral, that is, seemingly insufficient, and even irrelevant, to understanding the mind and brain. This perspective manifests itself in widespread neuroscience chauvinism. For instance, decision making is no longer by animals, but commonly described as executed by neurons. Indeed, whole experiments are often described as oriented to neurons: "Neurons were presented with images," as if the animal housing those neurons is a mere inconvenience (see Popivanov, Schyns, & Vogels, 2016 for a random example). I believe that this reductionist approach is fundamentally flawed—neurons work in concert with other neurons, across many different brain areas. Assuming that a process such as decision making can be understood by looking at single neurons, or even their ensembles, is like trying to understand why people in Australia drive on the left side of the road by studying their DNA. It is the wrong level for investigating many pressing questions in neuroscience in general, and in cognitive science in particular. Even if we could measure all the neurons in the brain at once, without an incisive behavioral paradigm we would not be able to answer meaningful scientific questions. Indeed, in my own research, behavior has told me more about the brain than neuroscientific data have, and in teaching the course "Introduction to Cognitive Neuroscience" I often struggle to infuse the class with neuroscience rather than focus only on clever experimental designs and the resultant revealing behaviors (Coltheart, 2006).

An extremely sobering example is the nematode *Caenorhabditis elegans* whose behavior we are still far from being able to predict, despite it having only 302 neurons and probably the best characterized nervous system in the universe (Bargmann & Marder, 2013). The importance of studies of *C. elegans* notwithstanding, this suggests that when trying to understand the human brain, and in particular high-order cognitive functions, a focus on neurons is perhaps leading us down a rabbit hole. In this side of the debate, I will argue that the "hierarchy" of cognitive science and neuroscience research, in which neural measurements are seen as basic and fundamental, and behavior is an optional component that cannot stand on its own, should be reversed, with behavioral research restored to its historical place of primacy and necessity.

One might counter that 50–60 years ago we did not have better options—non-invasive magnetic resonance imaging of humans had not yet been developed, and optogenetic manipulations were only a dream—and this is the main reason why the majority of discoveries that have withstood the winds of time are attributed to behavioral research. However, even in this era of impressive methods for recording and manipulating the brain, the results seem to mostly verify what we have already learned about cognitive function from behavior, rather than lead to novel insight. Of course, neuroscientific measurements (or brain perturbation studies) are necessary for answering questions about localization—*where* in the brain different (behaviorally identified) cognitive functions are implemented. But at the end of the day, behavior is still the standard by which theories about the *roles* and *computations* carried out by different neural structures are tested (even for theories that do not seek to model behavior directly; see, for example, Wang, 2008). If we ask ourselves candidly "what has neuroscience taught us about cognition that we did not already know from behavior?" we realize that unfortunately the answer is "not very much at all." Therefore, in the name of expediency, we have a moral obligation to our subjects of study, our taxpayer funders, and the patients waiting for cures, to reverse the trend that suggests that we should dispense with behavioral research and focus only on neural mechanisms.

1. What Has Behavior Taught Us About the Brain?

Long before the advent of neuroscience methods such as optogenetics and DREADDs (Designer Receptors Exclusively Activated by Designer Drugs), ingenious behavioral experiments made great strides in identifying the latent cognitive processes that underlie behavior—the ultimate output of the brain. For instance, even in low-level perception, arguably the area of neuroscience furthest from behavioral output, scientists were able to infer that color vision is due to three cones—and were even able to estimate the cones' wavelength sensitivity—from psychophysics experiments of color matching and chromatic adaptation and studies in color-blind observers (e.g., Stiles, 1959; Stiles & Burch, 1959). This occurred decades before physiologists were able to patch and measure the wavelength response properties of individual cones and verify the original behavior-based predictions (Baylor, Nunn, & Schnapf, 1987). At the other end of the spectrum, the elusive domains of high-level cognition and cognitive control, ideas about the role and information content of attentional signals feeding back into low-level perceptual processing areas were also derived from simple, but revealing, behavioral paradigms such as visual search and pop-out (Hochstein & Ahissar, 2002), and from experiments delineating the extent to which practice effects generalize across simple visual tasks (Ahissar & Hochstein, 1993, 1997). These two examples from the domain of vision are representative of others in all levels of research into the workings of the brain and the mind. In this section, I describe in detail a small selection of results from behavioral studies that illuminate brain processes, each going beyond what would have been achievable even using unrealistic, whole-brain, single-cell resolution neural measurements.

I will start with my personal favorite: A rat is trained to run in a T-shaped maze, from the base of the T to the right arm, in order to obtain food. At the right/left decision point, the rat may be making the choice to go right based on internal (egocentric) cues, i.e., turning right relative to its own body, or based on external (allocentric) cues, turning towards a certain location in space, for example to the side of the room that has a window. Now, imagine you could record activity from anywhere in the rat's brain during this experiment: How would you determine which of the two strategies the rat was using? It is not clear what you should look for in your recordings. In particular, the fact that the external cues of the window and blinds are represented in some brain area does not mean the rat is using these cues to guide its actions. Arguably, even if you had recordings of every neuron in the brain during this task, it would be extremely difficult to answer the simple question regarding the rat's strategy. And the problem is not that recordings are correlational rather than causal. Perturbing a brain area and seeing an effect on behavior will also not reveal how the animal was making its decision when there was no perturbation. By manipulating the brain we can find out what brain areas *can* affect behavior on this task, but not what brain areas *do* affect behavior as the rat is making its right turn.

Instead, what Packard and McGaugh (1996) did was simple and cheap: They turned the maze around, so that if the base of the T was pointing to the south, now it was pointing to the north. They then set the rat at the base of the T, and observed its behavior. If the rat continued to turn right, that would indicate an egocentric strategy. If the rat turned left, it must be following the peripheral external cues, and turning to the same side of the room. The latter is indeed what happened for most of the rats after eight days of training, but not after 16 days of training, suggesting that decision-making strategies change with extended training. This elegant behavioral manipulation thus informed us of a neural strategy, suggested representations that are necessary for executing it (the rat must be representing the external cues), and allowed for a variety of follow-up experiments that would delineate the conditions under which rodents choose to use an allocentric rather than egocentric strategy. Understanding these conditions then led to computational models that specified the types of data that the animal must learn and use in each condition, and the computations that may support transitioning from one strategy to another (Daw, Niv, & Dayan, 2005). Indeed, neuroscientific experiments involving lesioning or inactivating brain areas continued to reveal some computational principles of this decision-making process (see Section 2); however, the initial findings were due to cheap, but extremely clever, behavioral experiments.

In fact, the most fundamental observation about learning—that it proceeds through error-correction—was based on behavioral findings. In particular, if you train an animal that some stimulus, say, a flashing light, predicts the availability of food, the animal will learn this relationship, as can be measured from the animal's salivation response once the light begins flashing. Similarly, if a tone is paired consistently with food, the animal will begin to salivate to the tone. However, in a series of carefully controlled experiments in the 1960s, Leon Kamin showed that if the light is first paired with food until learning asymptotes, and only *then* the tone is added to the light (still with food following presentation of both), the

animal will not learn to salivate to the tone (Kamin, 1968). This phenomenon of "blocking" showed that presenting two stimuli in a predictable temporal relationship is not sufficient to engender learning. Instead, learning requires a "prediction error"—the motivationally significant outcome (i.e., the food) must not already be fully predicted (in this case, by the light). Based on these and other seminal behavioral findings, Robert Rescorla and Alan Wagner proposed a computational model of learning based on prediction errors (Rescorla & Wagner, 1972) that is, to this day almost 50 years later, the most influential account of trial-and-error learning in animals and humans. Together with later computational models of reinforcement learning showing how one can use prediction errors to optimally learn the sum of future rewards predicated on a certain state or stimulus (Barto, Sutton, & Anderson, 1983; Sutton, 1988), and recordings of activity of dopaminergic neurons while monkeys learned to associate cues with motivationally significant outcomes (Ljungberg, Apicella, & Schultz, 1992), these findings led to the influential identification of phasic dopaminergic signals with a reward prediction-error signal (Barto, 1995; Montague, Dayan, Person, & Sejnowski, 1995; Montague, Dayan, & Sejnowski, 1996), widely heralded as the poster-child of computational neuroscience. This success story—a normative computational theory that explains and predicts key neural signals with precision (e.g., Waelti, Dickinson, & Schultz, 2001; Tobler, Dickinson, & Schultz, 2003), drilling in all the way from Marr's (1982) computational level, through an algorithmic solution, and to its neural implementation (Niv, 2009; Niv & Langdon, 2016)—began with behavioral findings (Sutton & Barto, 1990) and is a prime example of behavior and computation informing our understanding of neural mechanisms, and not the other way around.

In the domain of memory processes, the phenomenon of "retrieval-induced forgetting" shows that, when trying to recall an item, memory traces that are similar enough to compete for recollection but eventually lose the competition are subsequently weakened (Anderson, 2003). For example, imagine learning the word-pairs FRUIT-PEAR, FRUIT-KIWI, and FRUIT-APPLE. In a later rehearsal phase, you are requested to complete the stem FRUIT-PE__ (presumably with the word "pear"). Since "apple" is the quintessential fruit, and moreover it bears some semantic similarity to a pear (round-ish, palm-sized, tree-growing fruit that is sometimes green), it may come to mind as you recall the word "pear" and compete for that recollection. Behavioral findings have shown that if the competition is sufficiently strong, the memory trace for "apple" is subsequently weakened, such that you are less likely to later remember the pair FRUIT-APPLE in a recall test, or to complete the stem FRUIT-A____ with "apple"; moreover, you may even tend to not remember the word "apple" in other pairs (Anderson, Bjork, & Bjork, 1994). This effect is item-specific (it does not affect "kiwi," a fruit that was less likely to come to mind and compete with "pear"); it depends on competition for retrieval (just rehearsing the pair FRUIT-PEAR does not weaken "apple"), and its boundaries in terms of dependence on practice, memory strength, etc., have been extensively characterized (Anderson & Spellman, 1995). The behavioral phenomenon of retrieval-induced forgetting imposes constraints on the implementation of memory systems, suggesting specific forms for networks of memories, and how and when they are updated. In particular, it has suggested that the

function of oscillating inhibition seen in cortical semantic memory networks and hippocampal attractor networks of episodic memory may be to strengthen weak memories and punish competitors (Norman, Newman, Detre, & Polyn, 2006; Norman, Newman, & Detre, 2007)—an insight that would be much harder to glean from neural recordings.

Another example is research into the format and structure of working memory in general (Miller, 1956) and visual working memory in particular (Brady, Konkle, & Alvarez, 2011). Here, too, progress on the question of whether working memory takes the form of a limited number of discrete slots or rather is limited due to a shared resource—questions that seem most reliably about the neuroscientific implementation of a computational component of perception—has relied on behavioral psychophysics research as well as computational modeling (Bays & Husain, 2008; van den Berg, Shin, Chou, George, & Ma, 2012; Keshvari, van den Berg, & Ma, 2013; van den Berg, Awh, & Ma, 2014; Brady & Alvarez, 2015), with neuroscientific evidence coming later, characteristically confirming rather than revealing the phenomenon (Ma, Husain, & Bays, 2014).

Finally, detailed hypotheses about how attention is deployed and controlled have also been inspired by and tested in behavioral data. In particular, recent work suggests that even as we are attending to an object, or a location in space, our attention spotlight is not stationary—it is moving around briefly, scanning the environment and returning back to the focus, at a frequency of 8 scans per second (Fiebelkorn, Saalmann, & Kastner, 2013). This "blinking of the attentional spotlight" may be related to other 8Hz (theta rhythm) processes characteristic of environmental sampling. It also constrains the search for brain areas that are responsible for deployment of attention. In particular, further neural data from non-human primates performing the same behavioral task implicate the lateral intraparietal cortex and frontal eye fields, two prominent parts of the classic frontoparietal attention network, in both maintaining attention and suppressing shifts away, and periodically shifting attention to rapidly scan the environment (Fiebelkorn, Pinsk, & Kastner, 2018). However, importantly, Fiebelkorn et al.'s (2013) findings that initially revealed the existence of this fascinating cognitive process did not involve any neural measurements. Moreover, the Fiebelkorn et al. (2013) findings could hardly have been predicted even knowing that other environmental sampling processes operate at 8Hz. Indeed, it seems that no amount of neural data on attention could have led us to posit this pattern of behavior with confidence, whereas a single behavioral study was sufficient to establish it.

2. What Have We Learned from the Brain That Behavior Had Not Already Taught Us?

In the companion chapter, Cushman argues that studying the brain itself is useful for cognitive research. This is clearly not in dispute. The question I ask is whether neuroscientific measurements are *necessary* for understanding cognition, whether they are primary and their import overwhelms that of behavioral findings, and whether studies that do not employ neuroscientific measurements should be categorized as "only" behavioral. Although neuroscience and behavior need not be

at two opposing ends of acceptable research—in fact, they are most effective when done in tandem—unfortunately, in recent years, the rise of neuroscience has been accompanied by a disdain for behavioral research to the point of rejecting its usefulness altogether. Meetings (e.g., Computational Systems Neuroscience), journals (e.g., *The Journal of Neuroscience*—a society journal that supposedly represents the whole field) and, perhaps most importantly, funders (e.g., the National Institute of Mental Health) all share the assumption that behavior, on its own, cannot lead to valid neuroscientific findings, and that understanding the mind is irrelevant to understanding the brain. In the name of trying to understand the brain (and not only cognitive or psychological processes), they reject out of hand studies that employ behavioral measurements alone.

However, when trying to enumerate what we have learned about cognition from neural measurements or perturbations that we did not already know from studies of behavior, we discover that the list is embarrassingly short. Of course, beginning from the Hodgkin–Huxley model (Hodgkin & Huxley, 1952)—a fundamental breakthrough in understanding how neurons fire action potentials and communicate that is of course not attributable to any behavioral findings—and continuing with many investigations of anatomy, physiology, and recordings of single neurons while animals were performing different tasks, we have learned a lot about the workings of neurons in different areas of the brain. However, the breakdown of different functions and cognitive processes into modules that can be expected to be realized independently in neural hardware, and that we should therefore be looking for in different areas, is predominantly thanks to behavioral studies of healthy individuals and those suffering from brain damage. Similarly, the most interesting findings from neural recordings or perturbations are when those are coupled to incisive, hypothesis-driven behavioral experimental designs. It seems that neuroscience alone is not nearly as revealing as a combination of a telling behavioral design, neural recordings or perturbation of neural function (e.g., through a lesion, inactivation, or stimulation), and a computational model that states the hypotheses to be tested precisely. For this combination to reach its peak performance, work on developing and testing behavioral paradigms, as well as computational models, is as important as work on developing new neuroscience methods. My argument is therefore that pure behavioral research is not only critical for understanding the mind—it is also the cornerstone of understanding how cognitive processes are implemented in the brain.

From the perspective of cognitive science in particular, the goal is to understand how the mind processes information—that is, what representations and computations are realized by mental activity. Cushman lists some classic questions in cognitive science: visual object recognition, language production, and theory of mind. It is therefore relevant to ask: What have we learned about these specific questions from neuroscientific research? Discovering, for instance, that face perception is special was initially done using behavior (Kanwisher & Yovel, 2006)—e.g., through comparing perception of right-side-up and upside-down faces (R. K. Yin, 1969; Diamond & Carey, 1986)—long before face cells were discovered in the inferotemporal cortex (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Of course, determining whether there are specialized modules for processing faces

versus other objects is aided by neuroscientific data (Kanwisher, McDermott, & Chun, 1997; Grill-Spector, Knouf, & Kanwisher, 2004; Yovel & Kanwisher, 2004); however, even those distinctions are often easier to make behaviorally rather than neurally (Leder & Carbon, 2006; Robbins & McKone, 2007). Regarding the mechanisms of language production, we already know much about the processes by which words are retrieved from a lexicon and strung together into syntactically correct sentences, and how meaning is extracted from sentences by a listener, mostly from work in linguistics that pre-dated the functional neuroimaging boom. Studies of communication in the human brain are only now starting to catch up (Dikker, Silbert, Hasson, & Zevin, 2014; Silbert, Honey, Simony, Poeppel, & Hasson, 2014; Simony et al., 2016; Yeshurun, Nguyen, & Hasson, 2017). For theory of mind, the concept, its breadth, and its development were all successfully studied without looking into the brain (e.g., Wellman, Cross, & Watson, 2001). Neural studies have revealed potential brain areas underlying the cognitive processes involved in theory of mind (Saxe & Kanwisher, 2003; Saxe, Moran, Scholz, & Gabrieli, 2006; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010), but have not explained the cognitive process of theory of mind as a series of representations and computations over these. Even domain specificity and dissociations between false beliefs and false photos were seen behaviorally through developmental work before their respective neural counterparts were pinpointed (as reviewed in Saxe, Carey, & Kanwisher, 2004).

Obviously, questions such as "what area of the brain is involved in process X?" can only be asked at a neural level. However, my claim is that understanding what process X is, and what computations it embodies, is rarely done at the neural level alone, if at all. Some exceptions can be found in the domain of perception; one can convincingly argue that measuring the receptive fields of neurons at different levels of the hierarchy of visual processing areas has helped explain how visual processing proceeds from building blocks to percepts (although we are still far from understanding even this basic process, and recently computational models, not neuroscientific data, seem to be providing most of the breakthroughs). But is perception the "highest" cognitive function for which neuroscience can inform cognitive science?

Although few and far between, there are some instances where neuroscientific research has led to insights about higher cognitive functions that would perhaps not be available otherwise. This has been mostly by way of lesion studies (or, in animals, reversible inactivations) that illuminated the separable components of cognitive processes once thought to be unitary. In the domain of memory research, for instance, the groundbreaking case of Henry Molaison (more commonly known as patient H. M.) first suggested a separation between episodic long-term memory and other forms of memory and learning, revolutionizing the study of memory and the hippocampus, and spurring the field of cognitive neuropsychology (Scoville & Milner, 1957; Augustinack et al., 2014). Notably, the findings relied on testing with appropriate behavioral tasks that demonstrated the different dissociations, e.g., between forming new episodic memories and learning new skills (Milner, 2005). Similarly, while behavioral studies such as Packard & McGaugh's (1996, discussed above) suggested that animals transition between

decision-making strategies as they become more familiar with a task, lesion studies in rodents paired with theoretically sophisticated behavioral paradigms revealed that both strategies—goal-directed decision making that relies computationally on forward planning, and habitual responses that lean exclusively on past experience (Dickinson, 1985; Dickinson & Balleine, 2002; Daw et al., 2005)—are learned in parallel and, in principle, available for use at any given time (Killcross & Coutureau, 2003; Yin, Knowlton, & Balleine, 2004, 2005; Balleine, 2005)

Other questions regarding apparent (or true) equivalences between categories of behaviors or responses still await resolution. For instance, is not getting an expected reward equivalent to losing money? Behavior suggests that these two situations are similar, but whether they truly are equivalent is a question best answered at the implementational level. Another question of this same flavor is: Do we have two antagonistic motivational systems, an appetitive one and an aversive one (Konorski, 1967), or rather is motivation controlled by one system with two poles? This class of questions—which fall squarely within cognitive science and are not merely implementational—can possibly only be answered at a neural level, albeit coupled with a suitable behavioral task. Indeed, these are prime examples for which a generic behavioral test (e.g., extinction learning or conditioned place preference) will not be nearly sufficient, and the neural findings must rely on a clever behavioral task specifically tailored to the question at hand.

I therefore want to make clear that I am not arguing for the futility of neuroscientific research in cognitive science. If anything, I am calling for a true merger between psychological cognitive science (ultimately interested in understanding behavior) and neuroscientific cognitive science (ultimately interested in explaining the brain). The study of behavior cannot afford to ignore such an important source of information—why would we be interested in measurements such as response times and eye movements but *not* be interested in accompanying neural measurements? And similarly, the study of the brain must rely on an understanding of behavior if it is to have any chance of making rapid progress.

3. Clever Experimental Design Can Make Up for Correlative Measures

One last point I would like to make, while on this soapbox, regards another strong bias in neuroscience, the preference for causal manipulations rather than correlational measurements. This predisposition suggests that since an fMRI signal is correlational, it is by design inferior to a technique that manipulates the neural hardware, for instance, optogenetically. This bias similarly renders behavioral measures in humans absent a brain manipulation (for instance, using transcranial magnetic stimulation, or due to a lesion) automatically inferior. However, I believe that clever behavioral designs have two advantages: They allow us to sidestep the causation/correlation pitfall and they help us utilize resources judiciously.

We can start by taking a cue from how the brain goes about making sense of the world. Indeed, we often infer causal constructs (e.g., in perception: "What

I am looking at is a table surrounded by four chairs") from what can be construed as noisy, low-dimensional, correlational measurements (e.g., the two-dimensional images that fall on my retina) together with prior knowledge about the generative process (that is, how different causes will manifest in such measurements). So, inferring causality from convergent correlational measurements is not an ultimate sin. The brain also utilizes perturbations—we can move our head to get a different image on our retina if an obscured object seems particularly ambiguous. Or we can walk over to the object in question and move it to verify that all the parts that we thought were connected really do belong to the object. Inferences are therefore reliant on correlational methods very commonly, and on costly perturbation methods in extreme cases. We can construe scientific inquiry similarly: In trying to infer the causal structures of the world around us (Why does application of a painful stimulus generate long-lasting fear responses? Why are humans so prone to assuming, even on first glance, that some people are more intelligent than others?), we should combine both correlational measurements and causal perturbations.

The next step is to realize that causal manipulations in neuroscience are not limited to silencing or activating a set of neurons, or lesioning a part of the brain. A behavioral task that requires a cognitive process can effectively constitute a causal manipulation, turning a neural process (and its underlying hardware) "on" and "off" through changes in task demands. As an example, imagine the N-back task, in which a participant views a series of stimuli and has to respond whenever the current stimulus is identical to the one viewed N trials back (with N being set in advance, for instance, to 2). This task requires constant maintenance of working memory, introducing and removing stimuli from the "recent N items" set, and comparing the current item to the contents of working memory. Changes in task demands could turn working memory mechanisms on or off, depending on the specifics: Setting N to 1 would place minimal requirements on updating of working memory as the task of identifying repetitions can be solved even at the level of iconic perceptual working memory; or the task can be changed to "respond to upside-down objects," which does not require working memory at all. Of course, task performance requires more than one cognitive process (e.g., in the N-back task: reading, memorizing, comparing the current stimulus to the content of memory, deciding on a response), so a clever, well-controlled experimental design is necessary to single out one function and not others. But this is exactly what the rich legacy of experiments in cognitive science has taught us to do well.

As illustrated by the examples in Section 1, this means that, using behavior alone, we can investigate even the neural implementation of working memory. For instance, by assessing the capacity of working memory for colors, orientations, and their conjunction, Luck and Vogel (1997) showed that visual working memory was stored at the level of whole objects (i.e., in higher-order visual perception areas) and not at the level of individual features. More recently, Katus and Eimer (2018) used the quintessential "causal" behavioral manipulation of dual-task requirements: They presented participants with visual and tactile stimuli, separately varying the number of items in each modality, and testing for working memory of only one modality in each trial. The two tasks therefore effectively activated the

neural mechanisms responsible for visual and tactile working memory, allowing one to assess whether these mechanisms are shared or separate. The behavioral results—no reduction in working memory accuracy for one modality with the increased demands on the other modality—suggest that independent memory storages exist for each modality, and that capacity constraints on working memory do not result from a shared higher-level control process (Cowan, 2010).

Thus, a relatively simple experiment can answer a question about the segregation of different types of information in working memory, without ever measuring neural signals. The equivalent neural perturbation experiment is dauntingly difficult and less incisive: One would presumably perturb the activity of a brain area (e.g., by TMS) and look for effects on working memory performance. However, finding that the perturbation does not affect task performance does not mean the brain area is not involved in the task (there could be redundancy in the system), and finding that performance is decreased does not specifically imply deterioration of working memory (here, again, one has to design careful control conditions to rule out other cognitive processes that are involved in the task). Finally, it is not clear that perturbation or neural measurement would tell us conclusively whether tactile and visual working memory rely on shared or separate neural substrates. This is therefore an example where a purely behavioral experiment can answer a neural implementation question more readily than neuroscientific measurements and perturbations can.

To be sure, here again I am not suggesting that neuroscientific measurements are irrelevant, or that there is no conceivable neuroscientific experiment that would answer the above question about working memory. What I am arguing is that the opposite is not true: It is not the case that pure behavioral experiments, using clever experimental designs, and behavioral output such as choices and reaction times, cannot possibly answer a question about neural mechanism. To the contrary, the latter may be better suited to answering that question with little time and effort. Considering behavioral measurements as inferior *a priori* is detrimental to neuroscientific progress.

4. Summary

I have illustrated above the huge asymmetry between behavioral research and neuroscientific investigation in cognitive science: Behavioral work contributes more to our understanding of the brain than neuroscience has contributed to understanding the mind. Interestingly, this is similar to the asymmetry between machine learning and computational modeling, which have contributed much more to neuroscience than they have learned from the brain. Perplexingly, in both cases, funding decisions—which are essentially prioritization exercises—suggest the exact opposite. This reversal may be fueled by several misconceptions. One is a prevalent illusion that neuroscientific data are in some sense "objective," whereas using a (computational) theory to interpret (behavioral) data is more "subjective" and less scientific. Another, not unrelated, widespread misconception is that behavior is already understood or not interesting, whereas the "real" questions are ones in neuroscience. But we are still far from explaining behavior (Rescorla, 1988) and closing shop in all

departments of psychological science. A question remains, then: What is the best way to make progress in understanding behavior? Although behavior is generated by the brain, I have argued that it is not the case that understanding the brain is the best way to understand behavior. Behavior and neural data are informative about different things: If you are interested in behavior (as many of us are, in psychology and cognitive science), then you should study behavior.

This observation suggests that we need a prioritization of questions, not of techniques. Once the questions of interest are clearly defined—for instance, "What is a reliable diagnostic phenotype of bipolar disorder?" (to take a translational example that seems most readily yielding to neural biomarkers)—we can consider what the best techniques are to answer the question expeditiously. Which brings me to the final illusion: That because behavioral work is sometimes easier (and very often cheaper), its findings are somehow worth less than findings from neuroscientific research. This is a common illusion in economics—people (and animals) value an expensive good more than they do a cheap one, and prefer a reward that they have earned through more effort to one achieved more easily (Aronson & Mills, 1959; Clement, Feltus, Kaiser, & Zentall, 2000; Plassmann, O'Doherty, Shiv, & Rangel, 2008). But this preference is irrational. Sunk costs do not actually make something more valuable, only more … expensive. Do we not, therefore, have a moral obligation to do the behavioral experiments—those that may lead to faster answers at a lower cost—first? Concretely, if a series of behavioral tasks, coupled with computational modeling of the cognitive processes underlying behavior on these tasks, can quantify the aspects of bipolar disorder that fall outside the norm defined by healthy patients, should we not pursue this route before subjecting patients to expensive neuroimaging scans?

In basic cognitive science research, the role of neuroscience has been suggested to be that of "constraining theories of representation and computation" (Cushman, p. 122, this volume). It is wholly unclear, however, whether the most effective constraints come from neuroscientific data or from behavioral data. In particular, the brain is complex and forgiving—it represents many quantities that may be ancillary to a specific function, and can solve a specific problem through several, often redundant mechanisms. As a result, neural data are often only weakly constraining, if at all. It may be our limitation as researchers (working with the cognitive abilities that our neurons afford us), but historically we have distilled more insight into cognitive processes from contrasting behavior in well-crafted experiments than we have from measuring or perturbing the brain. If we are to accelerate the progress of cognitive science and move even closer to understanding the human mind, we therefore should restore behavior to its rightful place as the base of cognitive science (Griffiths, 2015) on which other findings and types of data can rely, but without which one cannot build a thesis.

Acknowledgments

I am grateful to Daniel Bennett, Mingbo Cai, Nicole Drummond, Sarah Dubrow, Valkyrie Felso, Andra Geana, Gecia Hermsdorff, Angela Langdon, Angela Radulescu, Nina Rouhani, Nicolas Schuck, Melissa Sharpe, Yeon Soon Shin,

Olga Lositsky, Diksha Gupta, and Tyler Bonnen for helpful discussions in a lab meeting long ago (and some more recently) that contributed critical examples and ideas to this chapter, and to Daniel Bennett, Nathaniel Daw, Tom Griffiths, Nina Rouhani, Geoff Schoenbaum, Yavin Shaham, Melissa Sharpe, and Fiery Cushman for encouraging and insightful comments on an early draft.

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, 90, 5718–22.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, 387, 401–6. doi: 10.1038/387401a0

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, 49(4), 415–45.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 20, 1063–87.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*, 102, 68–100.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, 59, 177–81.

Augustinack, J. C., van der Kouwe, A. J. W., Salat, D. H., Benner, T., Stevens, A. A., Annese, J., et al., Corkin, S. (2014). H.m.'s contributions to neuroscience: A review and autopsy studies. *Hippocampus*, 24, 1267–86. doi: 10.1002/hipo.22354

Balleine, B. W. (2005). Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior*, 86(5), 717–30. doi: 10.1016/j.physbeh.2005.08.061

Bargmann, C. I., & Marder, E. (2013). From the connectome to brain function. *Nature Methods*, 10, 483–90.

Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (p. 215–32). Cambridge, MA: MIT Press.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, 13, 834–46.

Baylor, D. A., Nunn, B. J., & Schnapf, J. L. (1987). Spectral sensitivity of cones of the monkey macaca fascicularis. *Journal of Physiology*, 390, 145–60.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science (New York, N.Y.)*, 321, 851–4. doi: 10.1126/science.1158023

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, 15, 6. doi: 10.1167/15.15.6

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, 11(4). doi: 10.1167/11.5.4

Clement, T. S., Feltus, J. R., Kaiser, D. H., & Zentall, T. R. (2000). "Work ethic" in pigeons: Reward value is directly related to the effort or time required to obtain the reward. *Psychonomic Bulletin & Review*, 7, 100–6.

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, 42, 323–31.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, 19, 51–7. doi: 10.1177/0963721409359277

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–11. Retrieved from http://dx.doi.org/10.1038/nn1560 doi: 10.1038/nn1560

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, 115, 107–17.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 308(1135), 67–78.

Dickinson, A., & Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Learning, Motivation and Emotion* (Vol. 3, pp. 497–533). New York: John Wiley & Sons.

Dikker, S., Silbert, L. J., Hasson, U., & Zevin, J. D. (2014). On the same wavelength: Predictable language enhances speaker–listener brain-to-brain synchrony in posterior superior temporal gyrus. *Journal of Neuroscience*, 34, 6267–72. doi: 10.1523/JNEUROSCI.3796-13.2014

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. *Neuron*, 99(4), 842–53.

Fiebelkorn, I. C., Saalmann, Y. B., & Kastner, S. (2013). Rhythmic sampling within and between objects despite sustained attention at a cued location. *Current Biology*, 23, 2553–8. doi: 10.1016/j.cub.2013.10.063

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, 135, 21–3. doi: 10.1016/j.cognition.2014.11.026

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, 7, 555–62. doi: 10.1038/nn1224

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36, 791–804.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117, 500–44.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior: Aversive Stimulation* (pp. 9–31). Miami, FL: University of Miami Press.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 17, 4302–11.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 361, 2109–28. doi: 10.1098/rstb.2006.1934

Katus, T., & Eimer, M. (2018). Independent attention mechanisms control the activation of tactile and visual working memory representations. *Journal of Cognitive Neuroscience*, 30(5), 644–55.

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology*, 9, e1002927. doi: 10.1371/journal.pcbi.1002927

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal conrtex of rats. *Cerebral Cortex*, 13, 400–8.

Konorski, J. (1967). *Integrative Activity of the Brain: An Interdisciplinary Approach*. Chicago, IL: University of Chicago Press.

Leder, H., & Carbon, C.-C. (2006). Face-specific configural processing of relational information. *British Journal of Psychology*, 97, 19–29. doi: 10.1348/000712605X54794

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67(1), 145–63.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17, 347–56. doi: 10.1038/nn.3655

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco, CA: Freeman & Co.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.

Milner, B. (2005). The medial temporal-lobe amnesic syndrome. *Psychiatric Clinics of North America*, 28, 599–611, 609. doi: 10.1016/j.psc.2005.06.002

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, 377, 725–8.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16(5), 1936–47.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3), 139–54.

Niv, Y., & Langdon, A. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, 11, 67–73. doi: http://dx.doi.org/10.1016/j.cobeha.2016.04.005

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation*, 18(7), 1577–610. doi: 10.1162/neco.2006.18.7.1577

Norman, K. A., Newman, E. L., & Detre, G. (2007). A neural network model of retrieval-induced forgetting. *Psychological Review*, 114, 887–953. doi: 10.1037/0033-295X.114.4.887

Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, 65, 65–72. doi: 10.1006/nlme.1996.0007

Plassmann, H., O'Doherty, J., Shiv, B., & Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 1050–4. doi: 10.1073/pnas.0706929105

Popivanov, I. D., Schyns, P. G., & Vogels, R. (2016). Stimulus features coded by single neurons of a macaque body category selective patch. *Proceedings of the National Academy of Sciences*, 113(17), E2450–E2459.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435, 1102–7. doi: 10.1038/nature03687

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist*, 43, 151–60.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, 103, 34–79. doi: 10.1016/j.cognition.2006.02.008

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87–124. doi: 10.1146/annurev.psych.55.090902.142044

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporoparietal junction in "theory of mind." *NeuroImage*, 19, 1835–42.

Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, 1, 229–34. doi: 10.1093/scan/nsl034

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20, 11–21.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences of the United States of America*, 111, E4687–E4696. doi: 10.1073/pnas.1323812111

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7, 12141. doi: 10.1038/ncomms12141

Stiles, W. S. (1959). Color vision: The approach through increment-threshold sensitivity. *Proceedings of the National Academy of Sciences*, 45(1), 100–14.

Stiles, W. S., & Burch, J. M. (1959). Npl colour-matching investigation: Final report. *Optica Acta: International Journal of Optics*, 6(1), 1–26.

Sutton, R. S. (1988). Learning to predict by the method of temporal difference. *Machine Learning*, 3, 9–44.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (pp. 497–537). Cambridge, MA: MIT Press.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, 23(32), 10402–10.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, 121, 124–49. doi: 10.1037/a0035234

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 8780–5. doi: 10.1073/pnas.1117465109

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412, 43–8.

Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. *Neuron*, 60, 215–34. doi: 10.1016/j.neuron.2008.09.034

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72, 655–84.

Yeshurun, Y., Nguyen, M., & Hasson, U. (2017). Amplification of local changes along the timescale processing hierarchy. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 9475–80. doi: 10.1073/pnas.1701652114

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of the dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19, 181–9.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *European Journal of Neuroscience*, 22(2), 505–12. doi: 10.1111/j.1460-9568.2005.04219.x

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1), 141.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 6753–8. doi: 10.1073/pnas.0914826107

Yovel, G., & Kanwisher, N. (2004). Face perception: Domain specific, not process specific. *Neuron*, 44, 889–98. doi: 10.1016/j.neuron.2004.11.018

Further Readings for Part IV

Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives on Psychological Science*, 4(4), 326–39.

Argues that complex psychological states do not play an explanatory role in cognitive science but are instead the proper objects of explanation, and that they are constructed out of more basic psychological phenomena that correspond to processes in the brain.

Bickle, J. (2015). Marr and Reductionism. *Topics in Cognitive Science*, 7(2), 299–311. https://doi.org/10.1111/tops.12134

Argues that neuroscientific methods can and do provide genuine explanations of cognition.

Churchland, P. S. (1980). A perspective on mind–brain research. *Journal of Philosophy*, 77(4), 185–207.

Churchland, P. S. (1989). *Neurophilosophy: Towards a unified science of the mind-brain*. Cambridge, MA: MIT Press.

A seminal book on the relationship between neuroscience and psychology whose Part II argues that understanding the mind requires understanding the brain via bottom-up neuroscientific methods. See also the earlier article published in the Journal of Philosophy.

Coltheart, M. (2006). Perhaps functional neuroimaging has not told us anything about the mind (so far). *Cortex*, 42(3), 422–7. https://doi.org/10.1016/S0010-9452(08)70374–5

Coltheart, M. (2013). How can functional neuroimaging inform cognitive theories? *Perspectives on Psychological Science*, 8(1), 98–103. https://doi.org/10.1177/1745691612469208

Argues that functional neuroimaging results do not adjudicate between competing psychological theories. See peer commentaries and author's reply in the Cortex publication.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2), 97–115. Retrieved from JSTOR.

Classic philosophical argument that implies, among other things, that multiple realizability severely limits the extent to which brain imaging can advance our understanding of cognition.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3), 480–90. https://doi.org/10.1016/j.neuron.2016.12.041

Argues for a pluralistic methodology on which behavioral studies provide more insight into cognitive processes than neuroscientific studies and should typically precede neuroscientific investigation.

Roskies, A. L. (2009). Brain–mind and structure–function relationships: A methodological response to Coltheart. *Philosophy of Science*, 76(5), 927–39. https://doi.org/10.1086/605815

Argues that the most plausible version of the challenge laid out in Coltheart (2006) has been met, and that the extent to which brain imaging can advance understanding of cognition depends on the nature of the mapping between brain structures and the cognitive function realized by those structures.

Roskies, A. L., & Craver, C. F. (2016). Introduction to the philosophy of neuroscience. *The Oxford Handbook of Philosophy of Science*. https://doi.org/10.1093/oxfordhb/9780199368815.013.40_update_001

A philosophical introduction to the philosophy of neuroscience that includes a discussion of brain imaging and its relevance to understanding human cognition.

Study Questions for Part IV

1) According to Cushman, why do neuroscientific theories do little to advance our understanding of the mind?
2) According to Cushman, how in principle could neuroscientific methods advance our understanding of the mind?
3) According to Cushman, how did single-unit recordings of retinal ganglion cells lead to significant breakthroughs in computational models of visual processing?
4) According to Niv, what are the benefits of privileging behavioral studies over neuroscientific studies? Why do behavioral studies provide these benefits?
5) According to Niv, under what circumstances can neuroscientific methods advance our understanding of the mind in the absence of behavioral methods? What distinguishes these circumstances from typical circumstances?
6) According to Niv, when is relying on neuroscientific methods and behavioral methods preferable to relying on behavioral methods alone? What distinguishes these circumstances from typical circumstances?

Part V

What Can Cognitive Science Teach Us About Ethics?



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

9 The Ethical Significance of Cognitive Science

Victor Kumar

The answers to many philosophical questions depend on contingent facts about the human mind. This is especially clear in the philosophy of mind and adjoining philosophical neighborhoods. Thus, philosophers draw on empirical research in cognitive science to understand the relationship between the mind and the brain, the nature of reason and emotion, and the ability to know the contents of other minds.

Yet, how could cognitive science be relevant to ethics, of all things? Ethics, we are commonly told, is normative rather than purely descriptive. Since empirical research in cognitive science answers only descriptive questions, some philosophers may protest that it cannot yield answers to normative questions in ethics. There is, in short, a well-known gap between "is" and "ought" and no amount of new scientific knowledge about what *is* can bridge the gap that exists between it and philosophical knowledge about what *ought* to be. It would seem, then, that even if its significance for other domains of philosophical inquiry is secure, cognitive science has no ethical significance.

This argument—against the possibility of ethics informed by cognitive science—fails on multiple fronts. And to see this one need not experience any doubt about the existence of the is–ought gap. First of all, many topics in *meta*ethics are descriptive rather than normative. Though we wish to know the extent of our obligations to others, we also wish to know what it is to form a moral judgment about an issue like this. First-order questions about the good and the right are normative, certainly, but second-order questions about the cognitive status of moral judgment and its link to motivation are plainly descriptive. Thus, even if cognitive science were forever closed off to first-order, normative inquiry, it might still provide an important source of evidence for second-order, descriptive inquiry in metaethics about the nature of moral judgment.

The ability of cognitive science to help answer second-order questions about moral judgment suggests a way in which the field can inform first-order inquiry too. Many of our moral beliefs are shaped by unconscious psychological processes. Some of these processes may be unsuited to confer justification. So, if empirical findings in cognitive science uncover the unconscious influences on certain moral beliefs, and if it turns out that these influences are epistemically disreputable, then the findings have a normative upshot: They *debunk* those moral beliefs. Even if cognitive science cannot reveal what we owe to others, it can tell us that

moral beliefs we hold about first-order issues like this are unjustified and should be abandoned.

Cognitive science is relevant to second-order questions in metaethics about the nature of moral judgment and the epistemic evaluation of moral beliefs. In the latter case metaethical conclusions have normative implications. But cognitive science informs ethical theory in a third way too, and here it has more direct normative significance. What we ought to do and who we should aspire to be are constrained by psychological feasibility. The significance of our psychological makeup lies in its import for *non-ideal theory*. In normative ethics, non-ideal theory eschews universal criteria of right and wrong and instead seeks more modest, but more useful, generalizations about moral improvement from our present circumstances. By revealing what is psychologically feasible, cognitive science can shed light on the very nature of moral progress.

My task in this chapter is to make good on the three promissory notes issued above. That is, I will show how cognitive science informs three topics of inquiry in ethics—the nature of moral judgment, debunking arguments, and non-ideal theory. My focus in this chapter is on metaethics and normative ethics, and thus I will set aside inquiry that fits more comfortably in the study of free will, moral responsibility, action theory, and applied ethics. In each section I will lay out the relevant philosophical subjects, explain how cognitive science can help address them, and then illustrate by discussing recent work in naturalistic ethics.

1. Moral Judgment

Some areas of metaethics have few obvious connections to cognitive science. The metaphysics of morality is, at best, only tangentially related to the science of moral cognition. However, it is far more plausible that cognitive science offers resources to that branch of metaethics that is concerned with the nature of moral judgment—where "moral judgment" is construed as a psychological state rather than a linguistic utterance. Philosophers working in this area of metaethics are interested primarily in two related topics: (1) whether moral judgment is a cognitive or non-cognitive psychological state and (2) whether moral judgment has a necessary or contingent link with motivation.

Some philosophers explicitly conceive of their target of study as the concept of moral judgment (Smith, 1994; Jackson, 1998). Others voice no such explicit commitment, but employ an *a priori* methodology that seems best suited to probe the concept. Yet, another possibility is that the target of study is not the concept of moral judgment, but moral judgment *itself*. Is this, however, a coherent thought? Is there any way to get a grip on moral judgment except as mediated by our concept of moral judgment? The answer, I believe, is yes.

Moral judgment is of interest because we are confronted by numerous examples of the phenomenon in everyday life. We make moral judgments about others and about ourselves, as do other people with whom we interact. Paradigm cases of moral judgment thus provide an alternative way of grasping the subject matter: We may study moral judgment itself by studying paradigm cases (cf. Kornblith, 2002). And if we want to know what the paradigm cases have uniquely in common, the

best way is via empirical study. Thus, if the target of study is moral judgment itself, we need not begin by analyzing our concept of moral judgment. The concept of moral judgment may inaccurately represent the real world phenomenon.

Metaethical naturalists need not start from scratch, methodologically, and can take a cue from decades of work in the philosophy of mind (Fodor, 1981, 1987; Griffiths, 1997; Schroeder, 2004; Prinz, 2004; Holton, 2009; Weiskopf, 2009). Traditionally, philosophers who study mental state categories have been interested in the concept of belief, or the concept of desire, etc. But nearly since the birth of cognitive science, it's been understood that ordinary concepts of mental state categories might be radically mistaken—hence the extensive debate about eliminativism regarding the posits of folk psychology (Churchland, 1981; Stich, 1983, 1996). If mental states themselves constitute the object of study, what we should do is empirically study them. To take this approach is to treat mental state categories (defeasibly) as *natural psychological kinds* and to adopt the methodological approach suited to investigation of putative natural kinds (Kumar, 2016a). As philosophers of language suggest, a natural kind term refers directly to the corresponding kind, usually via ostension to paradigm cases, even if the intension of the concept associated with a natural kind term picks out a different extension, and even if it denotes the empty set (Putnam, 1975; Kripke, 1980; see also Kumar, 2014).

Like other mental states, moral judgment has two ingredients: attitude and content (see Kumar, 2015). The bulk of attention in metaethics has been accorded to the attitude that is constitutive of moral judgment, rather than its content. The default view, perhaps, is that moral judgments are ordinary beliefs. To judge that disrespectful behavior is wrong is simply to believe that it is wrong. Historically, this cognitivist view of moral judgment has come under attack for both psychological and metaphysical reasons. On the one hand, moral judgments, unlike beliefs, seem to have an essential tie to motivation. Someone who says that it is wrong to cheat at cards, but goes on to do so without the slightest hesitation, seems not to be speaking sincerely (Smith, 1994). On the other hand, if moral judgments are beliefs, then their truth ostensibly requires the existence of moral properties, and the existence of moral properties has seemed dubious, as has our ability to gain epistemic access to them (Mackie, 1977). Better to save the legitimacy of moral thought—to avoid the threat of massive and widespread error—by understanding moral judgments as non-cognitive attitudes with no ontological purport.

A number of different non-cognitivist proposals have been put forward. The earliest view characterized moral judgments as emotions or states of approval and disapproval (Stevenson, 1937; Ayer, 1952). "Boo to cheating at cards." A more recent proposal casts moral judgments as non-cognitive attitudes of acceptance toward a norm (Gibbard, 1990). To judge that cheating at cards is wrong is, on this view, to accept a norm that forbids or discourages cheating at cards. Non-cognitivist views such as these are able to secure an essential link between moral judgment and motivation, and able too to avoid the embarrassing result that moral thought is perpetually and irredeemably in error.

Cognitivism, however, remains an attractive position. First of all, moral judgments seem to be truth-apt, and arguably only cognitive attitudes can be true

or false. Relatedly, non-cognitivists have a difficult time accounting for the validity of inferences in which one of the premises is a moral judgment. This is especially true in embedded contexts, where a moral proposition is embedded in a logically complex construction (Geach, 1965; Schroeder, 2008). Furthermore, even though moral properties are metaphysically and epistemically puzzling, it still seems as if we have moral knowledge, halting and fallible though it certainly is (Campbell, 2007). For instance, we know that Jim Crow laws in the US were morally wrong. Of course, knowledge seems to be a cognitive mental state. What's more, much progress has been made in understanding the existence of moral properties and the possibility of moral knowledge, both in naturalistic and non-naturalistic terms (see, e.g., Railton, 1986; Shafer-Landau, 2003). All of these considerations buttress cognitivist theories of moral judgment.

The philosophical literature on moral judgment is vast and the foregoing review of the cognitivism/non-cognitivism debate elides numerous complications, many of which involve each side attempting to accommodate considerations that have been thought to support the other side. Still, it is enough to gain an entry point. From the naturalistic perspective outlined above, if we want to know how psychological considerations—rather than, say, metaphysical considerations—bear on the debate, what we should do is set aside thought experiments and instead examine directly whether paradigm cases of moral judgment behave like cognitive or like non-cognitive states.

A further methodological hurdle must be cleared if empirical evidence can be infused into the cognitivism/non-cognitivism debate. The mind is complex, to say the least. How are we to individuate moral judgment from other mental states that are concomitants of it? For example, a wealth of evidence suggests that emotions are critical elements of moral cognition. But this does not show, by itself, that moral judgment itself consists—partly or wholly—in emotion. For all this empirical evidence says, emotions may be related to moral judgments only causally, rather than constitutively.

The solution to this problem is to examine not just what is present in paradigm cases of moral judgment, but also the *causal/explanatory role* of moral judgment in psychological generalizations (Kumar, 2016a). What we want to know, then, is what sort of causal role paradigm cases of moral judgment regularly play in our psychology and whether cognitive or non-cognitive states underwrite that role. To broaden our attention in this way is to adhere more closely to the methodology reflected in the scientific study of other natural kinds. Water is H_2O, and not some more complex formula that includes common solubles found in paradigm cases, because it is H_2O alone that underwrites the chemical explanations in which water figures.

Cognitive science informs metaethical investigation of moral judgment when the target of study is moral judgment itself, rather than the concept, because cognitive science facilitates empirical study of paradigm cases of moral judgment and their causal/explanatory role. Jesse Prinz and I are two philosophers who pursue this methodology (Prinz, 2007; Kumar, 2015, 2016a).

Prinz argues for a view that he calls "emotionism," not to be confused with "emotivism," the non-cognitivist view that moral judgments are emotions. Prinz

claims, rather, that moral judgments are beliefs. However, he argues that the moral concepts that are constituents of moral judgments are composed of emotions. More precisely, to have a moral concept is to have a disposition to undergo certain types of emotional episodes. To conceptualize an action as morally wrong is, roughly, to be disposed to experience guilt if one performs it and resentment if others do. Thus, Prinz's view is cognitivist but also says that emotions are constitutive of moral judgment.

In support of emotionism, Prinz (2007, pp. 21–3) cites empirical evidence that paradigm cases of moral judgment are invariably accompanied by emotion. However, as he makes clear, this evidence might show only that emotions are necessary causes of moral judgments, rather than constitutive of them (p. 23, p. 29). The clincher, according to Prinz, is evidence that emotions are not only necessary but also sufficient for moral judgment. In one study, participants judged quite innocent actions to be morally wrong when they were hypnotized to experience negative emotions in response to morally neutral words (Wheatley & Haidt, 2005). In a study of "moral dumbfounding," participants maintained moral judgments about emotionally evocative activities even when they could cite no reasons, suggesting that emotions are sufficient for moral judgment in the absence of states that might otherwise be proposed as constituents (Murphy et al., 2000). A number of criticisms might be raised against this argument, but the most pertinent is that it does not seem to adhere to the naturalistic standards outlined above that Prinz himself accepts. Emotions might be necessary and sufficient causes of moral judgment, for all the cited empirical evidence says, and yet they may still generate ordinary beliefs. The evidence Prinz cites does not distinguish cause from constitution.

Unlike Prinz, I defend a "hybrid" theory of moral judgment (Kumar, 2016a). This is a theory that marries cognitivism with non-cognitivism. On my hybrid view, moral judgment consists in a moral belief and a separable state of moral emotion (see also Campbell, 2007). To judge that someone acted wrongly is to believe that she acted wrongly and also feel, say, resentment toward her. Typical moral judgments have both elements, but atypical cases may have only one or the other.

I begin my case for a hybrid theory by considering two types of psychological explanations in which moral judgment figures. First, moral judgment explains reasoning not just about moral matters but also about many other domains. For example, moral judgments seem to influence mental state ascriptions in the Knobe effect (Knobe, 2010). Second, moral judgments also explain behavior—in particular, cooperative, uncooperative, and punitive behavior. Thus, in economic games, people make moral judgments about others and this influences whether they are likely to cooperate with fellow participants in prisoners' dilemmas, and whether they punish fellow participants in the ultimatum game (Fehr & Fischbacher, 2004). So, collectively, this evidence suggests that moral judgment plays the causal/explanatory roles of cognitive states *and* non-cognitive states.

I argue next that the non-cognitive state that is partly constitutive of moral judgment is moral emotion. My main source of evidence is a large body of work which suggests that two types of psychological processing are at play in moral

cognition, one generating beliefs and the other generating emotions (for review and discussion see Haidt, 2001; Greene, 2008; Campbell & Kumar, 2012). I also cite evidence suggesting that these two processes are tightly integrated in moral cognition (see especially Rozin, 1999), and thus that belief and emotion are bound together in a causal unity—a homeostatic property cluster (Boyd, 1991). This line of argument seems to adhere more closely than Prinz's to the naturalistic methodology suited to the investigation of natural kinds. The principal challenge for my hybrid theory, arguably, is whether other, competing accounts of moral judgment might also have the ability to underwrite its causal/explanatory role. This remains an open, empirical question.

I have focused on naturalistic investigation of the attitude that is constitutive of moral judgment. But another central question about moral judgment also seems capable of benefiting from a naturalized methodology. Is the link between moral judgment and motivation necessary or contingent? My theory and Prinz's theory suggest similar answers to this question. On Prinz's view, a disposition to experience motivating emotions is necessary, but this disposition is *ceteris paribus* and may be defeated. On my view, motivating emotions are present in typical, hybrid cases of moral judgment, but in some atypical cases only moral belief is present without any corresponding emotion or motivation. Thus, both views entail that motivation is linked constitutively with moral judgment, without being strictly necessary (see also Kumar, 2016b).

Application in metaethics of the naturalistic methodology so common in the philosophy of mind has only begun, but it seems to have promise. We may better understand what moral judgment is—cognitive or non-cognitive, necessarily or contingently motivational—if we draw on research in cognitive science that uncovers what is common to paradigm cases of moral judgment and supports their causal/explanatory role.

2. Debunking Arguments

If cognitive science can offer insight into the nature of moral judgment, the central methodological barrier it must cross is to successfully apply empirical criteria that distinguish cause from constitution. However, the causes of moral judgment are, indeed, philosophically significant in their own right. A leading view in epistemology is that the causes of a belief—both what initiate and what sustain it—determine its epistemic status. That is, a belief is justified or warranted insofar as it is caused by epistemically reputable processes. It would seem, then, that research in cognitive science on the causes of moral beliefs might help us epistemically evaluate them.

But what precisely counts as an epistemically reputable process for arriving at moral beliefs? Unfortunately, this question is fraught by disagreement at the level of first-order ethics. Utilitarianism suggests that a moral belief is justified if it is based on processes that are sensitive to aggregate happiness or wellbeing. Deontology suggests that a moral belief is justified if it is based on processes that apply legitimate principles about rights and duties. Naturally, other first-order ethical theories imply other, competing views about the processes sufficient to confer

justification on moral beliefs. The trouble, then, is that even if cognitive science isolates the processes influencing moral beliefs, there seems to be no uncontroversial way to infer that they are epistemically reputable.

Cognitive science may not be able to endow epistemic credit upon our moral beliefs. However, it *can* discredit them. Utilitarians and deontologists do not agree about whether a belief forming process attuned to happiness or wellbeing confers justification. But they must certainly agree that a wide range of other psychological processes are *epistemically defective*. For example, suppose we were to discover that certain moral beliefs are the products of wishful thinking or logical fallacies. In that case, no matter which first-order ethical theory is true, these beliefs would be *debunked*. Herein lies the possibility of ethical debunking arguments that rely on research in cognitive science (Nichols, 2014). If any such arguments are sound, we should conclude that the targeted moral beliefs are unjustified, that we should abandon them, and that any moral beliefs conflicting with them are now on firmer ground (see Kumar & May, 2019).

One source of skepticism about ethics informed by cognitive science is the existence of the is–ought gap. As we have seen, the nature of moral judgment is a descriptive topic, and thus metaethical naturalists who study it need not mind the is–ought gap. However, conclusions about the epistemic status of our moral beliefs are patently normative. Does this spell trouble? Not at all. Psychological debunking arguments in ethics require empirical premises derived from research on the causes of moral beliefs. But they also require one or more normative premises, i.e., that some psychological process is epistemically disreputable. Debunking arguments are effective only when the normative premises upon which they rely are more plausible than the moral beliefs they attempt to debunk (Kumar & Campbell, 2012). Moreover, then, debunking arguments explicitly depend on normative assumptions and do not even attempt to jump the is–ought gap. As we shall see over the course of this section, the critical challenge for debunking arguments is to articulate empirical premises that can bear the load placed upon them.

The most well-known type of debunking argument in ethics draws on evolutionary psychology (Ruse, 1986; Joyce, 2006; Street, 2006; Rosenberg, 2011). According to many scientists, human morality is fundamentally a biological and cultural adaptation for social creatures that faced a host of ecological challenges requiring cooperation and coordination. Philosophers argue that if morality is designed by natural selection, then it is unlikely to have hit upon objective moral truths. Evolved moral beliefs are constrained only by expediency, not by accuracy.

Evolutionary debunking arguments face a number of challenges. Most pressing, the required empirical premises are highly speculative. Thus, many evolutionary debunkers are eager to disavow any categorical commitment to evolutionary accounts of morality and resort to conclusions that are conditional in form: *If* morality is an adaptation, *then* moral beliefs are unjustified. Another problem is that it is not at all obvious that natural selection is an epistemically disreputable process (Kumar & May, 2019). To make this claim stick one needs to assume controversial premises in first-order ethics. For if, say, what grounds morality is desire-satisfaction (consequentialism) or social interaction upon reasonable terms

(contractualism), then it may well be that natural selection has been, as a contingent matter, sensitive to these grounds.

Much research in cognitive science is experimental, and some of it targets the psychological processes underlying moral cognition. This program of research is maturing and continues to face methodological challenges. Nonetheless, it is on firmer evidential ground than evolutionary psychology. Furthermore, it is uncontroversial that certain psychological processes are epistemically disreputable. Thus, it seems, research on the more proximal—rather than ultimate—causes of moral beliefs provides a more credible source of evidence upon which to found empirical debunking arguments.

To better understand how experimental research can debunk moral beliefs, and to understand the challenges that debunking arguments face, it will help to examine them more concretely. Notice, first, that these arguments fall on a spectrum from global to local, sweeping to selective. Some philosophers draw on cognitive science in an attempt to debunk all moral beliefs: We should withhold judgment not about some moral issues but all of them. Other philosophers attempt to debunk very specific moral beliefs, like, say, the doctrine of double effect. In what follows, I will illustrate how cognitive science can support debunking arguments in ethics, but I will also argue that these arguments can incorporate plausible empirical premises only if they are very selective; see Kumar & May (2019) for more detail.

Arguably, nearly all of our moral beliefs are based on intuitions about concrete cases. We believe that killing is wrong, in general, because particular instances of killing feel intuitively wrong. Walter Sinnott-Armstrong (2008) argues, however, that moral intuitions are distorted by cognitive biases. Sinnott-Armstrong cites evidence on framing effects: People who are presented with a moral dilemma make different judgments based on how the dilemma is framed. For example, intuitions are biased by the order in which dilemmas are presented, or by whether the dilemma highlights either lives lost or lives saved (e.g., Petrinovich and O'Neill, 1996). It is entirely uncontroversial that framing effects are an epistemically disreputable process. One should accept this premise no matter whether one subscribes to utilitarianism, deontology, or some other ethical theory. And so, if moral intuitions are based on framing effects, we have reason to abandon any moral beliefs nourished by intuition—which might include all of them.

This argument is ambitious—indeed, too ambitious for its own good. As Joshua May and I point out, framing effects account for only a small fraction of the influences on moral intuition (Kumar & May, 2019). If a more complete set of causes were offered, it would not be clear that all intuitive moral beliefs are based on disreputable processes. So, the empirical burden of this debunking argument has not been met, at least not if the aim is to debunk all intuitive moral beliefs. (To be fair, Sinnott-Armstrong's principal aim is not to debunk all moral beliefs, but to criticize an intuitionist moral epistemology.)

Joshua Greene (2008, 2014) develops a debunking argument with slightly narrower scope. His target is characteristically deontological moral intuitions, i.e., those associated with deontology. Greene argues that characteristically deontological intuitions are produced by heuristics and biases within a fast, automatic, unconscious psychological system. Furthermore, he argues that these heuristics and

biases often lead us astray. Greene himself has conducted studies which suggest that these intuitions can be influenced by morally irrelevant factors, like whether harm is inflicted by direct force rather than by remotely pressing a button. According to Greene, deontology is based on intuitions; these intuitions are unjustified; so too, then, is deontology. What's more, deontology conflicts with utilitarianism, and so if Greene is correct that we have reasons to distrust deontology, utilitarianism gains an edge.

Greene's debunking argument does not beg the question against deontology. All ethical theorists, deontologists included, should agree that whether harm is inflicted nearby or remotely is morally irrelevant; a belief-forming process that is sensitive to this difference does not confer justification. Nonetheless, Greene's debunking argument against deontology fails because his empirical premises do not stand up to scrutiny. It may be true that characteristically deontological intuitions are influenced by certain morally irrelevant factors. But these intuitions are also sensitive to a wide range of other factors that seem to be morally relevant: the intensity of the harm, whether it was caused intentionally, whether there are mitigating circumstances, etc. (see Kumar & May, 2019). Some of these factors may be morally irrelevant from the perspective of utilitarianism, but Greene cannot rely on a normative premise to this effect without introducing assumptions that are more controversial than the moral beliefs he targets (cf. Berker, 2009). Furthermore, new evidence in moral learning theory suggests that the fast, automatic, unconscious system underlying moral intuition can be flexibly attuned to changing material and social conditions (Railton, 2014; Kumar, 2017a). Once we understand more fully the causes of the relevant intuitions, then, it no longer seems as if they are epistemically disreputable.

Daniel Kelly (2011) also develops an ethical debunking argument and, like Sinnott-Armstrong and Greene, Kelly sets his sights on intuitive moral beliefs. However, Kelly's target is narrower still: all moral beliefs based on disgust. The consensus among cognitive scientists is that disgust is fundamentally an emotion designed to help us avoid the threat of disease and infection (Rozin et al., 2008; Tybur et al., 2013). Only later, as Kelly explains, was disgust "co-opted" for moral cognition. Nonetheless, by now many moral beliefs are based on feelings of disgust. This includes conservative repugnance toward homosexuality but also liberal abhorrence of GMO food products. Yet other moral beliefs, besides these, seem to be tied to disgust. Cheating, deception, and exploitation, for example, often elicit moral disgust (see Kumar, 2017b). The problem, according to Kelly, is that disgust is an unreliable psychological mechanism. It evolved to "play it safe rather than sorry," so to speak. That is, it was evolutionarily advantageous for disgust to be oversensitive: Better to be disgusted by innocuous objects than to fail to be disgusted by harmful objects. So, Kelly concludes, disgust is unreliable and should not be trusted in morality.

Kelly claims that disgust influences a range of moral beliefs. However, although available empirical evidence suggests that many types of moral violations elicit disgust (see Chapman & Anderson, 2013), the evidence also suggests that, by itself, disgust has little influence upon moral beliefs. May (2014) argues that the effect of disgust on moral belief is insubstantial and infrequent. A recent meta-analysis

by Justin Landy and Geoff Goodwin (2015) confirms the insubstantial effect of disgust on moral belief. Perhaps if disgust did substantially influence certain moral beliefs, we would have reason to abandon them (cf. Kumar, 2017b), but, as it stands, the empirical premise in Kelly's debunking argument does not seem to be supported by the available evidence.

Sinnott-Armstrong attempts to debunk intuitive moral beliefs by appeal to framing effects; Greene by appeal to simple heuristics and biases; Kelly by appeal to disgust. However, these arguments are unsound because they rest on dubious empirical premises. None of the aforementioned processes seem to exert a substantial influence on moral beliefs. The likely problem is that the target in each case is too broad. It is unlikely that cognitive science will provide a simple story about the influences on a broad and heterogeneous class of moral beliefs, such that the influences are epistemically defective. Perhaps, then, were we to focus on more narrow classes of moral beliefs, we would have some hope of finding disreputable influences.

Richmond Campbell and I offer a schema for selective debunking arguments (Kumar & Campbell, 2012; see also Kumar & May, 2019). The idea is to focus, narrowly, on a pair of divergent moral beliefs, and to investigate empirically why they diverge. Imagine two similar cases. We believe that action in one case is wrong, but that action in the other case is right. Imagine further that empirical research is deployed to study why we respond differently to the two cases—something for which an experimental approach is well-suited (Kumar & May, 2019). If it turns out that the *causally* relevant difference between the two cases is *morally* irrelevant, then it seems that our different beliefs about the cases are unjustified. We should abandon one or the other.

Let's consider a simple illustration. Judges are tasked with making parole decisions about prisoners and one of the constraints they should follow is to treat like cases alike. If John deserves parole, and Bill's case is not relevantly different from John's, then it is wrong not to offer parole to Bill. A recent study by Shai Danziger and colleagues (2011) finds that some judges do not treat like cases alike, and that their parole decisions are, unsettlingly, influenced by whether or not they are hungry. Prisoners who face the judges before lunch are much less likely to be granted parole than prisoners who face the judges after lunch. Clearly, whether or not the judges are hungry is an epistemically disreputable basis for beliefs about merited parole. So, this empirical evidence suggests that one set of beliefs must be abandoned—though by itself it does not tell us which. Either the late morning prisoners deserve a higher rate of parole, or the early afternoon prisoners deserve a lower rate of parole; see Kumar & May (2019) for further discussion.

No doubt there are other factors influencing the judges' beliefs about parole, but because we have focused narrowly on the different beliefs they have before and after lunch, the empirical premises in this debunking argument are plausible. Selective debunking arguments with a similar structure have recently been developed that target our reluctance to donate to charity (Kumar & Campbell, 2012) and the doctrine of double effect (Feltz & May, 2017).

Debunking arguments in ethics can rely on cognitive science to supply empirical premises about the causes of moral belief. The principal challenge for

debunking arguments is to identify credible empirical premises. These premises, I have suggested, will be more plausible if the arguments are highly selective—if they target narrow pairs of moral beliefs rather than large classes of beliefs. Of course, cognitive science cannot yield philosophical conclusions on its own. Philosophical work is needed to articulate plausible normative premises that do not beg important questions in first-order normative ethics. Still, empirical work in cognitive science that probes the unconscious influences on moral beliefs might lead us to discover that some cherished moral beliefs are based on epistemically disreputable grounds. Any such moral beliefs are unjustified and we should abandon them in the absence of better grounds. This is how cognitive science has not just metaethical but also normative significance.

3. Non-ideal Theory

So far, we've looked at two ways in which cognitive science can inform metaethical branches of moral philosophy. First, philosophers interested in moral judgment can understand its nature by attending to research on moral judgment and its causal/explanatory role. Second, philosophers interested in debunking moral beliefs can look to research on epistemically defective psychological processes, especially processes underlying certain narrow pairs of moral beliefs.

In the rest of this chapter we'll explore a way in which cognitive science informs *normative ethics*. As we have seen, if cognitive science is to furnish empirical premises that lead to normative conclusions, philosophy must do its part by supplying plausible normative premises. The critical normative premise that supports normative ethics of the sort developed here is a variant of *ought-implies-can*. Suppose you can fulfill a duty to a friend by performing one of two actions. If one action is more feasible, few would deny that you have more reason to perform it. Or suppose you can cultivate a virtue by one of two means. If one is more feasible, again it seems obvious that you have reason to take that means over the other.

Some of the most striking research in cognitive science over the past few decades examines the implicit biases that underlie certain forms of discriminatory behavior; see Brownstein (2015) for a review. This work is of obvious ethical importance in light of the severe negative consequences that implicit biases have on women, people of color, and other vulnerable and marginalized groups. Of particular importance is research that investigates how to reduce the influence of implicit biases (Lai et al., 2014). This research offers a guide about how best to fulfill our duty to eliminate or reduce unjust discrimination. Obviously, we should pursue the means that cognitive science suggests are most effective.

This example makes clear that research in cognitive science on psychological feasibility has the potential to inform ethical reasoning. But my concern in this chapter is with ethical *theory*, and so far it is not at all clear that research on things like implicit biases has any theoretical significance in ethics. The following picture seems to capture the division of labor between philosophy and science. *A priori* philosophy supplies theoretical claims about what we owe to others and what sorts of character traits are virtuous. Science, it seems, only offers guidance in the

application of these theoretical claims—no guidance in how to articulate theoretical claims in the first place.

I believe that cognitive science can also inform theory construction in normative ethics. But how? To make headway, we must lay out in detail a form of *non-ideal normative ethics*, beginning with the recent history of non-ideal theory in political philosophy.

A distinction between ideal and non-ideal theory finds its original formulation in the work of John Rawls (1971), who employs the distinction in relation to theories of justice. The aim of ideal theory is to understand the principles that structure an ideally just state, abstracting away from current injustices that our societies face. It doesn't matter whether people as they currently are would comply with these principles. For Rawls, we want to know what principles would structure an ideal society *if* people were to comply with them. The aim of non-ideal theory, by contrast, is to examine the current state of our societies, their most pressing injustices, and the series of steps that will lead incrementally to an ideally just state. As is evident, ideal theory is prior to non-ideal theory within Rawls' framework. We first must understand what an ideally just state looks like, and only then can we take the steps needed to appropriately address current injustices.

A very different understanding of non-ideal theory is found in the work of Amartya Sen (2009), Elizabeth Anderson (2010), and a number of other political philosophers. This view rejects the need for ideal theory at all. It seems unlikely that a single brilliant philosopher, or even a community of brilliant philosophers working together, will be able to understand what a perfectly just society looks like. Moreover, non-ideal theorists argue that we don't need to know what the ideal is in order to see what sorts of improvements are possible, what sorts of changes would enhance justice in our societies. The *theory* in this form of non-ideal theory consists in generalizations about progressive moral change. And the method for theory construction is broadly empiricist. What we should do is search for changes to current practices and to those in the recent past that seem to be progressive, and then draw general lessons about the nature of progressive change. The unit of analysis is not the ideally just state, nor the steps that lead to it, but types of changes that tend to lead to moral progress. For non-ideal theorists like Sen and Anderson, this is the most we can have and, fortunately, it is all that we need.

Non-ideal theory is typically pursued within political philosophy, but there seems to be a clear analogue in normative ethics. The aim of normative ethics, traditionally conceived, is to articulate standards of right action that provide a more-or-less unifying explanation of what right action consists in. Utilitarianism and deontology are classical theories in "ideal ethics." According to utilitarianism, an action is right because it maximizes aggregate happiness or wellbeing. According to Kantian deontology, an action is right because it follows from a rule that can be consistently universalized. Though competitors, each theory offers a universal account of what makes actions morally right.

Non-ideal ethics rejects not just the answers provided by utilitarianism and deontology, but also the very questions posed. It seems unlikely that a community of brilliant philosophers will be able to arrive at a grand unifying theory that explains what makes actions right in every conceivable circumstance. In fact,

particular moral insights are hard won by communities of diverse people facing moral problems together, not in the abstract but on the ground (Anderson, 2014, 2016). In addition, we don't need to have an ideal theory that specifies what one should do in every set of circumstances to make changes to our behavior, our character, and our institutions that are morally progressive. The aim of non-ideal ethics is not to provide a unifying explanation of right action. Rather, it is to draw generalizations about the sorts of changes that tend to promote moral progress. These generalizations will be local to the sorts of material and social conditions that people face now and have faced in the recent past, but for precisely that reason non-ideal theorists can more reasonably hope to arrive at justified conclusions, and to develop theories about progressive moral change that provide genuinely useful moral guidance.

Non-ideal ethics, unlike non-ideal political philosophy, is in the very early stages of development. Extended criticisms and defenses of the project, as such, have yet to be articulated. Nonetheless, concrete contributions to non-ideal ethics have begun to emerge, and to get a clearer sense of the project as a whole it will help to examine some of these contributions. I will not attempt to unite various pieces of work in non-ideal ethics under a single banner, eliding, for example, important contributions in feminist philosophy and philosophy of race. My interest here is specifically with work in non-ideal ethics that draws on cognitive science. I will limit myself, in the rest of this section, to non-ideal approaches to virtue and to moral reasoning.

John Doris (1998, 2002) has mounted an empirical critique of traditional Aristotelian virtue ethics (see also Harman, 1999). According to Doris, Aristotelian virtue ethics posits the existence of "robust" character traits, that is, traits undergirding behavioral dispositions that exhibit significant consistency across situational contexts. Bravery, for example, undergirds a disposition to face one's foes in spite of fear, whether one's foes are on the battlefield, in the boardroom, or in the press. Doris argues, however, that a wealth of research in social and personality psychology over several decades shows that people exhibit a remarkable lack of cross-situational consistency. To take just one example from a large body of research, psychologists find that whether or not people are willing to help another person in need often turns simply on whether they have recently found a coin in a telephone booth (Isen & Levin, 1972). The "situationist" critique of Aristotelian virtue ethics is, perhaps, best understood as a critique of a certain type of *ideal* virtue ethics. Aristotelianism articulates an account of moral character that, Doris argues, is not in fact instantiated in human beings and may be out of reach.

Peter Railton (2011), drawing on a similar body of evidence in social and personality psychology, offers a more constructive contribution to virtue ethics. He suggests that if we are to develop a psychologically adequate account of virtue, we should begin by examining the sorts of character traits that are commonly instantiated in human beings. First of all, agreeing with Doris and other situationists, Railton argues that we should turn away from robust character traits and instead focus on traits that are narrowly attuned to situational contexts. For example, Railton suggests that "justice" is not in fact a real character trait, and that some people have a strong capacity for institutional justice but a weak capacity for

interpersonal justice (2011, pp. 314–15). However, once we distinguish between the traits that underlie a capacity for institutional and interpersonal justice, we may understand how someone who possesses the former can learn to cultivate the latter too (pp. 319–20).

One character trait that seems to be genuinely instantiated in human beings is self-control—roughly, a capacity to delay gratification by exercising willpower and structuring one's perception of the appetitive temptations in one's environment. People vary in their capacities for self-control, and studies suggest that those high in self-control are more likely to achieve a number of distinct goods for themselves and for others (Mischel et al., 1992). Self-control is not, however, a robust character trait. One learns to exercise it in some contexts but not in others. According to Railton, empirical work by Peter Gollwitzer (2009, cited in Railton, 2011, p. 322) offers clues about how to develop situational self-control. Open-ended plans tend to be ineffective. Participants in Gollwitzer's studies are more likely to exert self-control by implementing "if-then" plans that are cued to particular situational contexts. If one is trying to lose weight, a plan simply not to eat too much food is of little help. It is more effective to form the plan that "*if* it is past 9:00 pm, *then* I will not eat." Thus, Gollwitzer's research seems to offer general lessons about how to cultivate self-control in a psychologically feasible way. More broadly, one can now see how psychological research on actual human character traits and their development could be used to develop a richer, non-ideal virtue ethics.

Let's turn now from character to moral reasoning. In non-ideal ethics the aim is not to arrive at a static criterion of right and wrong. Even if there were such a criterion, we cannot pretend to know what it is. We are even likely to be ignorant about what many of the right moral questions are, or how best to pose them. What's more likely to be within our grasp, however, are methods of improvement, ways of arriving at better guides for moral action, even if the very best guide forever eludes us. Railton (2011, 2014) and Kwame Appiah (2006) argue that unconscious processes of habituation are critical for moral improvement. Elizabeth Anderson (2014, 2016) argues that a certain type of social activism is an ineliminable tool for overcoming biases that afflict those in positions of power. She sees this type of activism as instrumental in the American abolitionist movement and suggests that it is likely to be useful in other, contemporary movements. However, another important tool for moral improvement is moral reasoning.

Rawls (1971) famously proposed that normatively ideal moral reasoning aims to arrive at wide reflective equilibrium (see also Daniels, 1979). We begin by bringing together our considered judgments about particular cases, plausible moral principles, and other normative and empirical theories about human beings and society. We then search for conflicts between these sets of claims. To arrive at reflective equilibrium we revise or modify claims in each of the sets, preserving those that seem the most plausible, until we achieve consistency. Thus characterized, wide reflective equilibrium presents itself as a supremely rational process, and so offers a seemingly apt tool for moral improvement.

The problem, however, is that while wide reflective equilibrium might be an apt tool for certain kinds of creatures, it seems to be of limited use to human

beings. First, bringing together *all* of our beliefs connected to morality is a herculean task, and figuring out which of these to revise is perhaps no less difficult. Second, except for certain single-minded moral philosophers like Bentham or Kant, nearly everyone is a *moral pluralist* (Ross, 1930). We do not think one moral principle reigns supreme. We accept a number of different moral norms and values that sometimes come into conflict. Unless we pretend that some ordinal ranking of these norms and values is possible, wide equilibrium seems to be a chimerical goal. Third and finally, empirical evidence suggests that people are often likely to accept principles not because they find them independently plausible but, rather, because they serve to *rationalize* the moral beliefs they hold about concrete cases (see Haidt, 2001). And so, any actual attempt to reach reflective equilibrium threatens to devolve into mere rationalization.

Richmond Campbell and I invite ethicists to consider another type of moral reasoning (Campbell & Kumar, 2012). "Consistency reasoning" consists, roughly, in treating like cases alike. It is typically a targeted process of social reasoning. For example, suppose that although I am an avid dog owner, I also eat meat. A vegetarian friend, however, presses me: What's the difference between factory farming and practices that I already consider abhorrent, like dog fighting? Faced with this challenge, and provided that I am disposed to trust my interlocutor, I should either decline to condemn dog fighting or, more likely, change my opinion about factory-farmed meat. Unless I can find some morally relevant difference between these two practices, I should treat like cases alike.

Campbell and I point out that consistency reasoning is a common, socially embedded mode of moral change, not just in philosophical debates but also in the law and in everyday social engagement. Furthermore, we argue that consistency reasoning is psychologically feasible because it is implemented by two preexisting psychological systems commonly described in so-called "dual process" models of moral cognition (2012, pp. 291–6). We suggest too that consistency reasoning has likely been an engine of progressive moral change, for example in attitudes toward homosexuality (2012, pp. 287; see also Kumar & Campbell, 2016). So, if we're looking for a tool for moral improvement, one that can be implemented by human beings and that has a good track record, consistency reasoning seems to offer more hope than wide reflective equilibrium. What the debate turns on is which type of moral reasoning is psychologically more feasible and which tends to lead more often to instances of progressive moral change.

Non-ideal ethics eschews traditional questions in moral philosophy about universal criteria for right action or about ideally virtuous agents. Rather, non-ideal ethics begins with our current behavior, character, and institutions, and then asks what sorts of recent and current changes are feasible and morally progressive. From this it draws generalizations about the nature of progressive moral change. The target of theoretical analysis is not ideal behavior, or ideal character, or ideal institutions, but incremental moral progress in each of these moral domains. The philosophical foundations of non-ideal ethics are just beginning to be unearthed. Still, the project has obvious appeal. A number of different social science disciplines are potentially relevant to non-ideal ethics, including anthropology and

history. But cognitive science also seems to offer guidance about the psychological feasibility of various moral changes. This is how empirical research into the human mind can advance theory construction in normative ethics.

4. Conclusion

Skepticism about naturalistic ethics may stem from the belief in a gap between is and ought. In fact, however, the is–ought gap does not support skepticism. Some topics in ethical theory are descriptive rather than normative. Ethicists who seek answers to questions about the nature of moral judgment should examine research in cognitive science. Furthermore, normative inquiry in ethical theory can also draw on cognitive science provided that normative premises are supplied. We can better assess our moral beliefs by examining their bases and subjecting them to normative scrutiny. Finally, we can also understand the nature of moral progress by studying the psychological feasibility of moral change, in light of a normative framework that privileges the non-ideal over the ideal. This may not be all that cognitive science offers to ethics, but it is quite enough to merit enthusiasm about this new form of philosophical naturalism.

References

Anderson, E. 2010. *The Imperative of Integration* (Princeton, NJ: Princeton University Press).

Anderson, E. 2014. The Quest for Free Labor: Pragmatism and Experiments in Emancipation. *The Amherst Lecture in Philosophy*, 9, 1–44.

Anderson, E. 2016. The Social Epistemology of Morality: Learning from the Forgotten History of the Abolition of Slavery. In M. Brady & M. Fricker (Eds.), *The Epistemic Life of Groups* (Oxford: Oxford University Press), 75–94.

Appiah, K. 2006. *Cosmopolitanism: Ethics in a World of Strangers* (New York: Norton).

Ayer, A. J. 1952. *Language, Truth and Logic* (Mineola, NY: Dover Publications).

Berker, S. 2009. The Normative Insignificance of Neuroscience. *Philosophy and Public Affairs*, 37, 293–329.

Boyd, R. 1991. Realism, Anti-foundationalism and the Enthusiasm for Natural Kinds. *Philosophical Studies*, 61, 127–48.

Brownstein, M. 2015. Implicit Bias. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (spring 2015 edition). http://plato.stanford.edu/archives/ spr2015/entries/ implicit-bias

Campbell, R. 2007. What Is Moral Judgment? *Journal of Philosophy*, 104, 321–49.

Campbell, R., & Kumar, V. 2012. Moral Reasoning on the Ground. *Ethics*, 122, 273–312.

Chapman, H., & Anderson, A. 2013. Things Rank and Gross in Nature: A Review and Synthesis of Moral Disgust. *Psychological Bulletin*, 139, 300–27.

Churchland, P. 1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy*, 78, 67–90.

Daniels, N. 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy*, 76, 256–82.

Danziger, S., Levav, J., & Avnaim-Pesso, L. 2011. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences*, 108, 6889–92.

Doris, J. 1998. Persons, Situations, and Virtue Ethics. *Noûs*, 32, 504–30.

Doris, J. 2002. *Lack of Character* (Cambridge: Cambridge University Press).

Fehr, E., & Fischbacher, U. 2004. Social Norms and Human Cooperation. *Trends in Cognitive Science*, 8, 185–90.

Feltz, A., and May, J. 2017. The Means/Side-Effect Distinction in Moral Cognition: A Meta-Analysis. *Cognition*, 166, 314–327.

Fodor, J. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science* (Cambridge, MA: MIT Press).

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

Geach, P.T. 1965. Assertion. *Philosophical Review*, 74, 449–65.

Gibbard, A. 1990. *Wise Choices, Apt Feelings* (Cambridge, MA: Harvard University Press).

Gollwitzer, P. 2009. Self-Control by If-Then Planning. Presentation to the Research Seminar in Group Dynamics, Institute for Social Research, University of Michigan (March 16).

Greene, J. 2008. The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders and Development*, vol. 3 (Cambridge, MA: MIT Press), 35–80.

Greene, J. 2014. Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, 124, 695–726.

Griffiths, P. 1997. *What Emotions Really Are: The Problem of Psychological Categories* (Chicago, IL: University of Chicago Press).

Haidt, J. 2001. The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108, 814–34.

Harman, G. 1999. Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society*, 109, 316–31.

Holton, R. 2009. *Willing, Wanting, Waiting* (Oxford: Oxford University Press).

Isen, A., & Levin, P. 1972. Effect of Feeling Good on Helping: Cookies and Kindness. *Journal of Personality and Social Psychology*, 21, 364–88.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press).

Joyce, R. 2006. *The Evolution of Morality* (Cambridge, MA: MIT Press).

Kelly, D. 2011. *Yuck! The Nature and Moral Significance of Disgust* (Cambridge, MA: MIT Press).

Knobe, J. 2010. Person as Scientist, Person as Moralist. *Behavioral and Brain Sciences*, 33, 315–65.

Kornblith, H. 2002. *Knowledge and Its Place in Nature* (Oxford: Oxford University Press).

Kripke, S. 1980. *Naming and Necessity* (Cambridge, MA: Harvard University Press).

Kumar, V. 2014. "Knowledge" as a Natural Kind Term. *Synthese*, 191: 439–57.

Kumar, V. 2015. Moral Judgment as a Natural Kind. *Philosophical Studies*, 172: 2887–910.

Kumar, V. 2016a. The Empirical Identity of Moral Judgment. *Philosophical Quarterly*, 66, 783–804.

Kumar, V. 2016b. Psychopathy and Internalism. *Canadian Journal of Philosophy*, 46, 318–45.

Kumar, V. 2017a. Moral Vindications. *Cognition*, 167, 124–34.

Kumar, V. 2017b. Foul Behavior. *Philosophers' Imprint*, 17(15), 1–17.

Kumar, V., & Campbell, R. 2012. On the Normative Significance of Experimental Moral Psychology. *Philosophical Psychology*, 25, 311–30.

Kumar, V., & Campbell, R. 2016. Honor and Moral Revolution. *Ethical Theory and Moral Practice*, 19, 147–59.

Kumar, V., & May, J. 2019. How to Debunk Moral Beliefs. In J. Suikkanen & Antti Kauppinen (Eds.), *Methodology and Moral Philosophy* (New York: Routledge), 25–48.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., & Nosek, B. A. 2014. Reducing Implicit Racial Preferences: A Comparative Investigation of 17 Interventions. *Journal of Experimental Psychology: General*, 143, 1765–85.

Landy, J. F., & Goodwin, G. P. 2015. Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence. *Perspectives on Psychological Science*, 10, 518–36.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong* (London: Penguin).

May, J. 2014. Does Disgust Influence Moral Judgment? *Australasian Journal of Philosophy*, 92, 125–41.

Mischel, W., Shoda, Y., & Rodriguez, M. 1992. Delay of Gratification in Children. In G. Lowenstein & J. Elster (Eds.), *Choice Over Time* (New York: Russell Sage Foundation), 147–64.

Murphy, S., Haidt, J., & Bjorklund, F. 2000. Moral Dumbfounding: When Intuition Finds No Reason. Unpublished manuscript, Department of Philosophy, University of Virginia.

Nichols, S. 2014. Process Debunking and Ethics. *Ethics*, 124, 727–49.

Petrinovich, L., & O'Neill, P. 1996. Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology*, 17, 145–71.

Prinz, J. 2004. *Gut Reactions: A Perceptual Theory of Emotions* (Oxford: Oxford University Press).

Prinz, J. 2007. *The Emotional Construction of Morals* (Oxford: Oxford University Press).

Putnam, H. 1975. The Meaning of "Meaning." In K Gunderson (Ed.), *Language, Mind and Knowledge* (Minneapolis: University of Minnesota Press), 131–93.

Railton, P. 1986. Moral Realism. *Philosophical Review*, 95, 163–207.

Railton, P. 2011. Two Cheers for Virtue: Or, Might Virtue Be Habit Forming? In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, vol. 1 (Oxford: Oxford University Press), 295–330.

Railton, P. 2014. The Affective Dog and Its Rational Tale: Intuition and Attunement. *Ethics*, 124, 813–59.

Rawls, J. 1971. *A Theory of Justice* (Cambridge, MA: Harvard University Press).

Rosenberg, A. 2011. *The Atheist's Guide to Reality: Enjoying Life without Illusions* (New York: Norton).

Ross, W. D. 1930. *The Right and the Good* (Oxford: Oxford University Press).

Rozin, P. 1999. The Process of Moralization. *Psychological Science*, 10, 218–21.

Rozin, P., Haidt, J., & McCauley, C. 2008. "Disgust." In M. Lewis, J. Haviland-Jones, & L. Barrett (Eds.), *Handbook of Emotions* (New York: Guilford Press), 757–76.

Ruse, M. 1986. *Taking Darwin Seriously* (Oxford: Blackwell).

Schroeder, M. 2008. *Being For* (Oxford: Oxford University Press).

Schroeder, T. 2004. *Three Faces of Desire* (Oxford: Oxford University Press).

Sen, A. 2009. *The Idea of Justice* (Cambridge, MA: Harvard University Press).

Shafer-Landau, R. 2003. *Moral Realism* (Oxford: Oxford University Press).

Sinnott-Armstrong, W. 2008. Framing Moral Intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology*, Vol. 2 (Cambridge, MA: MIT Press).

Smith, M. 1994. *The Moral Problem* (Malden, MA: Blackwell).

Stevenson, C. 1937. The Emotive Meaning of Ethical Terms. *Mind*, 46, 14–31.

Stich, S. 1983. *From Folk Psychology to Cognitive Science* (Cambridge, MA: MIT Press).

Stich, S. 1996. *Deconstructing the Mind* (Oxford: Oxford University Press).

Street, S. 2006. A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, 127, 109–66.

Tybur, J., Lieberman, D., Kurzban, R., & DeScioli, P. 2013. "Disgust: Evolved Function and Structure." *Psychological Review*, 120, 65–84.

Weiskopf, D. 2009. The Plurality of Concepts. *Synthese*, 169, 145–73.

Wheatley, T., & Haidt, J. 2005. Hypnotically Induced Disgust Makes Moral Judgments More Severe. *Psychological Science*, 16, 780–4.

10 Putting the "Social" Back in Social Psychology

Colin Klein

1. Introduction: Two Views of Social cognition

I agree with a lot of what Victor Kumar says. Like Kumar, I agree that cognitive science can inform ethical theorizing. I think Kumar has identified several plausible routes by which cognitive science can do so. Abstracting from the details, anyone interested in the ethical life ought to care about what we *can* do and what we *typically* do when we deliberate about moral questions. The fact that our moral intuitions might be shaped by something other than the moral facts ought to trouble moral theorists (Greene, 2008). While it's possible to make that argument badly (as Berker, 2009 shows), it seems to me that Kumar has avoided the obvious pitfalls. I enthusiastically welcome Kumar's call for a move away from thinking of moral judgment as a capacity which manifests in a single discrete episode, and towards thinking of it as a dynamic, interactive, and ongoing process embedded in social contexts. Most of our important moral reasoning is done in extended conversation with others, and even solitary reasoning often embodies a dialectic structure (Stein, 2005; Laden, 2010; Campbell and Kumar, 2012).

Insofar as Kumar and I disagree, it's over how to interpret the empirical facts. Very roughly, I think that Kumar doesn't pay *enough* attention to the effects of social context. In that regard he is in good company. Insofar as we disagree, it is a symptom of a broader debate between two different models in social psychology. Kumar sides with one; I'm going to spell out the other, and show its consequences.

Here is a concrete example. Among folks of our education and class, it is vanishingly rare that one hears racist statements openly made in public life. This is (I'm told) quite a shift over a few decades back. Yet widespread racial disadvantage obviously still exists. What explains this mismatch between public avowals and private actions?

One explanation, which has captured the popular and philosophical imagination (Greenwald et al., 1998; Gendler, 2011; Strohminger et al., 2014), is that racism still exists in an *unconscious* form. It manifests in faint reactions, imperceptibly skewed judgments, the subtle recoil. The underlying cognitive states may not even be worthy of the name *beliefs*, given that they are implicit and associative (Gendler, 2008, Levy, 2014). While any particular implicit attitude may have a minuscule effect, implicit racism is so widespread and so insidious that it has profound consequences in the aggregate.

Here is a second explanation. Many people are still pretty racist. They know they are racist (or, at least, their racist beliefs are ordinary, perfectly accessible beliefs, whether or not they identify them *as* racist). They know you have to keep that sort of thing to yourself, though. People get very mad about open racism (even people who are actually racist themselves). The social costs just aren't worth it. *Quiet* discrimination persists.

These possibilities aren't strictly exclusive, but they point to two very different models of how one ought to understand social cognition. Following Fiske and Taylor (2013; also Fiske, 1993a), I'll call this the difference between thinking of humans as *Activated Actors* versus as *Motivated Tacticians*.

Activated Actors are pushovers, even despite their best intentions. As Fiske and Taylor put it, social environments "rapidly cue perceivers' social concepts, without awareness, and almost inevitably cue associated cognitions, evaluations, affect, motivation, and behavior." Hence people are hostage to "fast reactions, variously viewed as implicit, spontaneous, or automatic indicators of responses unconstrained by perceiver volition" (2013, p. 15). We may not *want* to be pushed around. But the only thing we can do is try to preempt or override our preponent reactions: Implicit responses are stubbornly resistant to top-down influence.

Motivated Tacticians, by contrast, are continually and strategically evaluating their social environment. Within a social context, agents are always "choosing among a number of possible strategies, depending on current goals" (Fiske, 1993a, p. 172). Explicit motivations thus make an important difference: Agents are constantly evaluating and re-evaluating social contexts based on their appraisals and their current goals.

One should understand these two pictures as broad poles within which different historical theories have fallen and between which social psychology has historically swung (Kihlstrom, 2004; Turiel, 2010; Fiske and Taylor, 2013). However, the difference between Activated Actors and Motivated Tacticians doesn't line up neatly with many distinctions that philosophers care about. Both emphasize the importance of unconscious processing: The Motivated Tactician also thinks that our appraisals of social situations and their demands can happen quickly and automatically (Bargh and Morsella, 2008). Both deny a classical picture of rationality on which we are objective and dispassionate agents. Within both accounts, there is disagreement about whether we are rational in some other useful sense.

Recall that both Kumar and I think that moral reasoning is a process, undertaken in an explicit or implicit social context. Moral reasoning is a species of social reasoning. Hence, on the Motivated Tactician model, we should also expect moral reasoning to be affected by social and tactical considerations. I think this possibility is both plausible and should make a difference to how we understand the empirical literature on social and moral cognition. So I will address several of the same phenomena that Kumar touches on, showing how the Motivated Tactician account gives a very different *philosophical* perspective on the same set of empirical phenomena.

Before I get down to it, a final bit about my own motivations. Activated Actor theories are interesting and worth taking seriously. But they have also had, in my opinion, the unfortunate effect of obscuring important social facts. Writing on the

research into implicit bias, for example, columnist David Brooks (2013) claimed: "Sometimes the behavioral research leads us to completely change how we think about an issue. For example, many of our anti-discrimination policies focus on finding the bad apples who are explicitly prejudiced. In fact, the serious discrimination is implicit, subtle and nearly universal."

Many would be surprised to hear that only implicit discrimination remains. Less than six months before Brooks wrote the above, the Associated Press released a survey of the US public on various racial attitudes. Among the highlights: 37% of respondents agreed with the claim that most blacks were lazy, 38% that they were irresponsible, and 42% that they were violent. Nor is this limited to the sort of *hoi polloi* who find internet comment sections inexplicably attractive. There is ample evidence of explicit racism even among those in power.[1] Further, there appears to have been relatively little movement in racial attitudes over the past decade, even among white liberal voters (Hutchings, 2009).

Contra Brooks, there appears to be plenty of work left to be done on the explicit front. The Activated Actor model is currently over-represented when philosophers talk about social psychology. I will argue that it is time to step back and take the "social" part of social psychology more seriously.

2. World Enough and Time

Both the Motivated Tactician and Activated Actor models are *cognitive* accounts. They are methodologically individualist, focusing on internal processes.[2] They differ in the sort of cognitive processes they appeal to, because they differ in the parameters that show up in their explanations. Very roughly, Activated Actor models are *very* individualist: The social world is an important source of input, but once we fix the proximal input we can largely bracket off the social facts in our explanations. Motivated Tacticians, by contrast, give a much greater explanatory weight to social factors, especially variable or long-term ones.

An example will illustrate the point. Both views think that ordinary humans are cognitively limited, especially in social situations. In most circumstances, we can't process everything with complete accuracy in real time. These limitations are a crucial part of our (well-documented) failures. Yet the *source* of those limitations differs on the two models.

Roughly speaking, the Activated Actor model continues a tradition on which observed limitations are fundamental architectural constraints of the cognitive system. Optimal behavior being impossible, we have a bias towards strategies and heuristics that rapidly get us to good enough answers (Simon, 1996). There is considerable debate *within* this tradition about whether these heuristics are innate or learned, adaptive or maladaptive, and so on (Kelman, 2013). On all versions, however, these limitations are fixed features of our cognitive apparatus.

By contrast, the Motivated Tactician picture is one on which limitations are (roughly speaking) the result of strategic interaction between numerous incompatible goals. Fiske reviews a number of different ways of cashing out these tradeoffs, describing the general problem as the tradeoff between "accuracy-oriented or open-minded motivation with confirmatory or close-minded motivation" (1993a,

p. 172). There is some sense in which (say) the tradeoff between accuracy and speed depends on our makeup (the gods need not compromise). Yet the details of these tradeoffs, Fiske emphasizes, are sensitive to what we want. As such, the parameters determining the costs of various tradeoffs are not fixed: They vary both with our appraisal of the social situation and with our motivations.[3]

The empirical difference is important. Consider explanations of gender stereotyping. On Activated Actor models, stereotypes are *triggered*. They enter into deliberation early, and unless they're suppressed they lead to unwanted consequences (Bargh & Chartrand, 1999). Stereotypes work like other forms of priming. Writing on stereotypes, for example, Bargh and Morsella claim that: "In contextual priming, the mere presence of certain events and people automatically activates our representations of them, and concomitantly, all of the internal information (goals, knowledge, affect) stored in those representations that is relevant to responding back" (2008, p. 76). That is an extremely general account—the same story is presumably true about non-social priming effects like (e.g.) semantic and associative priming (Stanovich & West, 1983; Seidenberg et al., 1984). Explanations of stereotypes don't actually require any special knowledge about the social milieu of the actor (other than the assumption that it was complex enough to transmit stereotypes).

Contrast this with Fiske's (1993a) Motivated Tactician account of how stereotypes enter into deliberation. Fiske notes that the prevalence and effect of stereotyping is strongly modulated by power: Powerful people tend to rely on stereotypes when dealing with subordinates, whereas subordinates stereotype the powerful far less often. She explains these facts in terms of mutually reinforcing processes of control and attention. The powerful have more demands on their time. It is comparatively more costly for them to pay attention to subordinates. Hence they must lean more heavily on fast strategies like stereotyping. Hence: "People in power stereotype in part because they do not need to pay attention, they cannot easily pay attention, and they may not be personally motivated to pay attention" (1993b, p. 621).

The opposite is true of the powerless. They have more motivation to attend to people who exert control over their lives. They have especially good reason to pay attention to the idiosyncrasies of those in power rather than treating them as mere instances of a type. Hence the asymmetry.

None of these factors, Fiske notes, need be obvious to the actors themselves. Further, stereotypes exert control not just directly, but by serving as an "implicit anchor" of which everyone in the culture is aware (1993b, p. 623). This means that stereotypes can exert power not (just) because they are widely believed, but because it is widely believed that they are widely believed. In an elegant study on the effects of beliefs about individual "brilliance" on participation in academic fields, Leslie et al. note that stereotyping may result in exclusion of women and African Americans both via overt prejudice and simply by discouraging "participation among members of groups that are currently stereotyped as not having this sort of brilliance" (2015, p. 265). That discouragement can, presumably, occur regardless of whether the stereotyped group actually *believes* the stereotype. Regardless of what you think of your own ability, it is unpleasant to contemplate

a future where you constantly have to prove yourself to people who wrongly assume that you're inadequate for the job.

Motivated Tactician models of stereotyping thus incorporate a wide variety of social facts in their explanations while remaining fully cognitive. Such accounts give correspondingly different recommendations about effective interventions. Fiske emphasizes the role of attention in modulating social expectations, for example, and the possibility that "if people pay more attention, at least some of them are less likely to stereotype" (1993b, p. 627). This also leads to differential imperatives depending on one's role in the organization: The powerful have both more need and more opportunity to change.

3. Terrible People, Terrible Reasoning

Fiske's account of stereotyping brings contingent social facts in to explain a phenomenon with obvious moral upshot. The shift to a Motivated Tactician model has a broader consequence as well: It shows the strong effects that context can have on the process of moral deliberation itself.

As I noted above, Kumar emphasizes a process that he calls "consistency reasoning" (see also Campbell and Kumar, 2012). Consistency reasoning consists in "treating like cases alike." Crucially, it is usually a *social* process that involves give-and-take with one's interlocutors. Kumar argues (plausibly, in my opinion) that consistency has been an engine for progressive social change. Yet I think consistency reasoning is far more complex in practice.

A fictional example will illustrate the point. Consider the following exchange from the television show *The Simpsons:*[4]

BART: Uh, say, are you guys crooks?
FAT TONY: Bart, is it wrong to steal a loaf of bread to feed your starving family?
BART: No.
TONY: Well, suppose you got a large starving family. Is it wrong to steal a truckload of bread to feed them?
BART: Uh uh.
TONY: And, what if your family don't like bread? They like … cigarettes?
BART: I guess that's okay.
TONY: Now, what if instead of giving them away, you sold them at a price that was practically giving them away. Would that be a crime, Bart?
BART: Hell no!

This seems to me to be a clear case of moral consistency reasoning. It is also a *failed* case. Bart has gotten to the wrong result in the wrong way. But why?

The Activated Actor model might cite ways in which Bart was *pushed* away from the obvious truth. Perhaps he relied on unreliable moral heuristics (Sunstein, 2005). Perhaps he put too much weight on the output of fast unconscious processing (Kahneman, 2011). Perhaps he was easily swayed by the emotional pull of Fat Tony's argument (Greene, 2008). Perhaps he fell prey to a morally troubling

in-group bias (Gino and Galinsky, 2012) or some other set of evolutionarily based biases (Wielenberg, 2010). Perhaps all these and more. But of course, then the knife twists: *We do the same thing*. Bart's failure differs from ours only in degree, not in kind. Even that is cold comfort, for most of these biases are hard to recognize from the inside. Bart's final confidence is as characteristic as his failures.

Yet it seems to me that this can't be all that's gone wrong. Bart also made a more basic mistake: He assumed that Fat Tony was *arguing in good faith*. It may be reasonable to extend that sort of charity to our interlocutors, at least initially. But it's also an assumption that we should be willing to abandon—especially if we hear a tortured set of analogies that leads to a suspiciously self-serving end. In such cases, sensitive moral agents should be inclined to rescind their initial charity, and stick to their initial judgments.

Moral deliberation is difficult precisely because we live in a world with a lot of bad people. Many people are self-serving liars, at least some of the time. Many people care more about making peace than about getting to the right moral answer. Most moral consistency reasoning also isn't done in the seminar room; it is done when one person has harmed another, and is trying to get off the hook. Sometimes people really care about the moral truth. Very often, they care about avoiding consequences.

That makes moral consistency reasoning very easy to *exploit*. This goes both ways. Most confessions to the police (at least from the actually guilty) are not the result of coercion. Interrogators rarely need the phone book under the hot lights. Instead, as Zulawski and Wicklander (2002) point out, the most effective interrogators simply help the accused *rationalize* what they did ("He treated you like a slave rather than a valued employee. The business was going under anyway, and no one would miss the money. So it's not like you hurt anyone, and you were only trying to pay the mortgage …"). The drive to find a narrative on which your actions come out permissible is a very strong one.

Or consider the incredible vitriol that hypocrisy attracts. At first pass, it's not clear why hypocrisy should be so troubling. There is a mismatch between the hypocrite's explicit moral judgments and their behavior. But as Wallace (2010) notes, mere inconsistency isn't normally treated as a terrible personal failing. Nor does it seem like personal inconsistency ought to undermine first-order moral judgments. Yet people often take hypocrisy as evidence of faulty reasoning. Why? The answer, I suggest, is that judgments of hypocrisy play a crucial functional role in regulating factors that would otherwise tend to undermine moral discourse. Wallace (2010) argues that hypocrisy judgments provide a check on inappropriate instances of the reactive attitudes: We blame the hypocrite for their own unjustified blame. Isserow and Klein (2017) further argue that the role of hypocrisy judgments is to undermine putative claims to *moral authority*. We often defer to others' moral judgments, particularly to those we think we can trust. But that puts one in a morally precarious situation, open to exploitation. Sensitivity to hypocrisy is necessary to regulate that vulnerability both for ourselves and for others.

Note, again, that while these are all cognitive facts, they depend on contingent social facts. It is only when we take into account the motivations of others that

we have to start worrying about lying and manipulation. It is only when we place some of the epistemic costs of moral deliberation onto perceived authorities that we have to start looking for hypocrites. And it is only when we try to accommodate the actual social facts about differential power structures that we have to think about potential asymmetries in attention, access, and control.

Does this show that Kumar's moral consistency reasoning is wrong, or limited in scope? I think not. Rather, it shows that it operates at two different levels. There is a global level, on which we want our moral principles and particular judgments to be in agreement. That is the sense of moral consistency reasoning we have in mind when we consider large-scale changes in our beliefs.

But consistency reasoning, I submit, also operates in *local* contexts. That is, we also want our beliefs about the local social situation to cohere. Much of our social and moral discourse is aimed at those local contexts, not at big-picture moral considerations.

The way we frame the situation, the judgments we make, and the justifications we offer to others should, ideally, come into equilibrium with our beliefs about the people we're with, the sort of arguments we expect them to find convincing, and the goals we think that moral conversation is trying to achieve. Sometimes the correct response to perceived inconsistency is to change our judgments or interpretation of a difficult case. Other times, we should change our beliefs about what our interlocutors think, or what the point of arguing over a moral point really is.

4. Experimental Situations Are Social Situations

Local consistency reasoning is interesting in its own right. It is also interesting because it gives the Motivated Tactician account an alternative take on many experiments in the cognitive science of moral judgment. As an example, consider Jonathan Haidt's well-known work on moral dumbfounding. Haidt and his collaborators constructed cases like the following:

> Julie and Mark, who are brother and sister, are traveling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other. So what do you think about this? Was it wrong for them to have sex?
>
> (Haidt et al., 2000, p. 20)

They find that subjects are often "dumbfounded": They judge that the case is wrong, they are unable to offer any good reasons for thinking so, and yet they stubbornly persist in their judgment (Haidt et al., 2000). This seems to be good evidence for a non-rational, non-deliberative, social intuitionist view of moral judgment (Haidt, 2001). While there are problems with the original (unpublished)

study, I take it that the *phenomenon* is familiar enough: We've all met people who simply dig in their heels and stick to their guns on difficult moral issues.

The real question is *why*. Here I think there is a further benefit to thinking about humans as Motivated Tacticians. For starters, we are deeply concerned not just with *what* people do but why they do it. That is especially important when we turn to imagining morally laden actions: As Kennett (2011) puts it, "In order for us to imaginatively enter another's point of view, we must be able to see the *point* of what they do" (2011, p. 182). An inability to do so engenders imaginative resistance. But as Kennett points out, the only reason why Julie and Mark want to sleep together is that it would be a "new experience" (2011, p. 185)—a difficult motivation to accept in the case of such a loaded taboo.

Indeed, Royzman et al. (2015), revisiting Haidt's study, found that most subjects simply didn't believe that Mark and Julie's actions would not be harmful. Liu and Ditto (2013) similarly found that moral evaluation had an effect on factual appraisal of dilemmas: Subjects were less willing to accept parts of the setup of moral dilemmas that conflicted with their overall moral judgment. Put another way: Even explicit instruction is one more, *non-privileged*, input into local consistency reasoning. It can be overridden if it's too implausible.

I think there is more to say. Given the right motivations, there can be a disconnect between the reasons you *have* and the reasons you *give*. That is true in any social situation. But there is good evidence that experimental situations—particularly ones that involve dialogue with experimenters—are social situations with unique demands. Milgram's (1963) work on experimental obedience is well-known, of course, but the effect can be more subtle. As Orne (1962) demonstrated, subjects will eagerly interpret experimental demands in a way that makes them meaningful, and will do so even if they are explicitly informed that they are engaged in a pointless task.

Further, experimental contexts are also *communicative* contexts. As such, subjects interpret and give information in conformance with Gricean conversational norms. Subjects tend to assume, in accordance with Gricean norms, that all of the information they are given is relevant.

This is a well-known confound in studies of heuristics and biases (Hilton, 1995; Kelman, 2013). Apparently irrational behavior can often be redeemed if we pay attention to conversational pragmatics. Suppose I ask my wife whether (hypothetically speaking) she would be mad if I slept with one of her friends. That might make her a little suspicious about my fidelity. Suppose I asked her whether (hypothetically speaking) she would be mad if I slept with her friend Natalie. She would likely become *very* suspicious. Ought we convict my wife of irrationality? Given that Natalie is one of her friends, the latter can't possibly be more likely than the former. But of course, the fact that I specifically mentioned *Natalie* demands an explanation—conversational norms require us to be as specific as necessary and no more. The most obvious explanation is that I have Natalie on my mind. Attention to conversational pragmatics thus redeems what looks like a simple failure to appreciate probabilities.

The same processes occur in empirical studies of moral reasoning. Subjects tend to give only information that they think will be relevant to the experimenter. In

a nice study of this effect, Norenzayan and Schwartz (1999) looked at subjects' causal attributions about the actions of a mass murderer. They found that subjects gave more situational explanations when told that they were talking with a social scientist, and more dispositional explanations when they thought they were talking to a personality psychologist.

In sum, the fact of being in an experimental situation can have non-negligible influences on how subjects behave. That effect, I argue, ultimately stems from the effect of local consistency reasoning. Subjects want their reaction to particular cases to be consistent with both their general beliefs about the world and their particular beliefs about the situation they are in.

With that in mind, here's an alternative interpretation of Haidt's experiment. Subjects faced with the Julie and Mark case are asked whether that action was wrong. They judge that it's incest, and incest is wrong. That's all pretty obvious; the experimenter has foregrounded the incest-relevant bits. Gricean norms thus make the stated question puzzling. Nor does it seem like they are being asked to justify the *principle* that incest is always wrong, as opposed to passing judgment on a particular case. The experimenter is a member of the subjects' moral community, and they have no special reason to doubt that the experimenter also thinks that incest is wrong. Nor does the experimenter seem seriously invested in changing the subjects' moral beliefs on the topic. Hence subjects can rely on what Tetlock (1992) calls the "acceptability heuristic," according to which least-effort solutions to a moral problem are probably the correct ones. Given this, subjects focus less on whether the action was *wrong* and more on whether Julie and Mark ought to be *blamed*—that is, on the question about their willingness to enforce judgments in complicated cases.

Once you get to that point, it might be quite reasonable to get hard-headed and to ignore putative excuses. In a discussion of the fundamental attribution error, Tetlock notes:

> One way of pressuring other people to behave is by indicating to them that one has a low tolerance for justifications or excuses and that one will treat their behavior as automatically diagnostic of underlying intentions and personality attributes … As Axelrod notes, people are not only expected to act in accord with prevailing norms, they are also expected to censure those who violate norms (a norm to enforce norms, or a metanorm). Insofar as accountable subjects feel that their moral mettle is being tested (their willingness to apply metanorms), they may be more motivated to hold others responsible and to reject situational explanations or excuses … The argument simply asserts that accountable subjects will tend to rely on the acceptability heuristic when they can infer the views of the prospective audience. In this case, subjects may assume that the prospective audience expects them to hold people responsible for deviant or untoward conduct.
>
> (Tetlock, 1992, p. 361)

It is not unusual that Julie and Mark have a good excuse for violating a moral norm. *Everyone* has an excuse. Good moral agents don't take those excuses at face

value, even if experimenters would like them to. In Royzman et al.'s (2015) study on dumbfounding, they found that the instance of true dumbfounding dropped to virtually zero when subjects were carefully asked about their commitments and which question they thought they were answering.

5. Moral Development

This leads to a final point. Explanations like Fiske's of stereotyping or mine of dumbfounding cite differential power relationships *between* groups. But of course there are also likely to be substantial and morally important differences *within* groups: differences of motivation, of personality, of inclination, and of ability. The Motivated Tactician account paints moral deliberation—particularly in social contexts—as relying on forms of social competence. People obviously differ in their social competence. One might wonder whether those differences also make a difference to judgments.

Activated Actor accounts have largely downplayed individual differences. They belong to an experimental tradition that seeks, in Bakan's (1967) terms, *general* facts that hold of all subjects, not just averages about aggregates. At best, individual differences are explained in terms of variation of these basic elements. Usually, "individual variation is a source of embarrassment to the experimenter," mere error variance that obscures the main effect of interest (Cronbach, 1957, p. 674). As Cronbach (1957) points out, there is a second tradition in psychology that is concerned with correlation and explanation of individual differences.[5]

One obvious place that individual differences become important is in the study of moral development. Following theorists like Piaget (1932) and Kohlberg (1969), studies of moral judgment have assumed that moral judgment proceeds in distinct stages rather than springing forth fully formed. Further, moral competence might be the sort of thing that takes time and effort to develop. In a famous passage from the *Analects* about his own moral development Confucius claimed that:

> At fifteen, I had my mind bent on learning. At thirty, I stood firm. At forty, I had no doubts. At fifty, I knew the decrees of Heaven. At sixty, my ear was an obedient organ for the reception of truth. At seventy, I could follow what my heart desired, without transgressing what was right.[6]

Moral development is a lifelong process, so one might reasonably expect people to be much better (or worse) at it at different points in their lives.

Further, as recent research has emphasized, development is a process that also occurs in dialogue between children and their peers, their parents, and other authorities. This dialogue, like other forms of cultural learning, occurs in carefully scaffolded environments (Sterelny, 2012). It also opens the possibility that children are far from passive recipients of moral dictates. Instead, as Turiel emphasizes:

> From the beginning, the developing person acts to give meaning to and to understand the world, and to make sense of social relationships. And with development, they come to evaluate the dictates of others, to distinguish

> between what they judge to be legitimate and illegitimate demands and directives, and to engage in relational reciprocal interpersonal interactions and communications. In the process of early development, they evaluate social relationships and consequently come to accept and reject parental directives, sometimes leading to conflicts with adults and peers. Children, like adults, also approach social relationships from the viewpoint of moral concerns, including concerns with the welfare of others and fairness.
>
> (2010, p. 559)

This long developmental process is thus a deeply *social* one, reflecting a lifelong engagement in both global and local consistency reasoning.

As such, the process of moral development—and our particular judgments in the short term—are likely to be deeply shaped by long-term facets of that society. There is an increasing body of research on the content and structure of this shaping process (Graham et al., 2011). This includes the effect of political (Graham et al., 2009) and religious (Hofmann et al., 2014) beliefs, and the broader cultural and socioeconomic background (Henrich et al., 2010).

Some have taken this variation to be an argument against the stronger forms of moral realism (Machery et al., 2005), citing precisely the sorts of considerations that Kumar addresses. Thinking of humans as Motivated Tacticians isn't a particularly happy place for the moral realist to rest: If moral reasoning is sensitive to tactical and strategic considerations, even our considered philosophical judgments may not rise above our time, place, and role.

I'd like to conclude on a slightly more optimistic note—one that I hope will be compatible with the points on which Kumar and I both agree. If we think of humans as Activated Actors, then variation is indeed problematic: People either get it right or they don't, and there's little more to say. If moral reasoning is social reasoning, however—if, as Fiske and Taylor put it, "[people's] social thinking is for their social doing" (2013, p. 15)—then the mere presence of variation doesn't mean that everyone is getting things wrong.

Social life is complex, motivations are diverse, and resources are widely distributed. We should recognize the profound influence that the social has on the personal, and from there on the moral. But once we do we can also think about how we might *construct* social contexts that are less susceptible to the most damaging of social effects. I've argued that some paradigmatic social psychology experiments failed to do so. But philosophers and cognitive scientists are well-poised to work out the intermingled effects of the social on the cognitive. What is required is turning our gaze outward, towards the ways in which social factors affect our moral judgments.[7]

Notes

1 As I write, the state of Pennsylvania is dealing with revelations that high-profile state officials, including judges and state prosecutors, regularly traded racist and sexist jokes over official email (*Washington Post*, Dec. 26, 2015, "Pornographic email scandal roils Pennsylvania politics"). There is also considerable debate about whether implicit biases

as measured by the IAT (Implicit Association Test) actually correlate with negative outcomes in a way that's not simply mediated by explicit bias; see Oswald et al. (2013) for a review.

2 There is also a rich tradition of methodologically holistic, social-functionalist approaches to morality—e.g., those drawing on Durkheim (1915) or the Marxist corpus. These have had far less influence on analytic philosophy, at least in recent years.

3 In this, the model fits well with more recent approaches to automaticity, which have stressed a continuous and strategic boundary between the automatic and the controlled (Bargh and Morsella, 2008; Kihlstrom, 2008). Further afield, there is an interesting parallel to recent work on cognitive control, which stresses that capacity limitations arise from the tradeoffs between efficiency and effectiveness in the control system, rather than being primitive architectural features (Botvinick and Cohen, 2014).

4 Episode 8F03, "Bart the Murderer." For those unfamiliar with the series: Bart is a young larrikin who has fallen in with mafia boss Fat Tony and his gang. Asked to hide a hijacked shipment of cigarettes, he has begun to worry about his new friends' motivations.

5 Cronbach associates this tradition most strongly with research on personality; while this is historically plausible, it is worth noting that there are personality researchers who focus on common individual processes (cf. Cervone, 2005). Cronbach also makes the fascinating suggestion that individual differences research—growing as it did out of research on IQ and industrial placement—tends towards the politically conservative, while the search for common features is liberal and democratic.

6 *Analects* Book 2, Chapter 4. Legge translation.

7 Thanks to Jessica Isserow and Victor Kumar for helpful comments on this draft, as well as audiences at the 2011 Pacific APA and the Social Psychology Brownbag Series at the University of Illinois at Chicago for helpful feedback on a distant ancestor.

References

Bakan, D. 1967. The General and the Aggregate. In his *On Method: Toward a Reconstruction of Psychological Investigation*, 34–6. San Francisco, CA: Jossey-Bass, Inc.

Bargh, J. A., & Chartrand, T. L. 1999. The unbearable automaticity of being. *American Psychologist* 54: 462–79.

Bargh, J. A., & Morsella, E. 2008. The unconscious mind. *Perspectives on psychological science* 3: 73–9.

Berker, S. 2009. The Normative Insignificance of Neuroscience. *Philosophy & Public Affairs* 37: 293–329.

Botvinick, M. M., & Cohen, J. D. 2014. The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science* 38: 1249–85.

Brooks, D. 2013. Beware Stubby Glasses. *New York Times*., Jan. 10, 2013.

Campbell, R., & Kumar, V. 2012. Moral Reasoning on the Ground. *Ethics* 122: 273–312.

Cervone, D. 2005. Personality architecture: Within-person structures and processes. *Annual Review of Psychology* 56: 423–52.

Cronbach, L. J. 1957. The two disciplines of scientific psychology. *American Psychologist* 12: 671–84.

Durkheim, E. 1915. *The Elementary Forms of Religious Life.* New York: Macmillan.

Fiske, S. T. 1993a. Social cognition and social perception. *Annual Review of Psychology* 44: 155–94.

Fiske, S. T. 1993b. Controlling other people: The impact of power on stereotyping. *American Psychologist* 48(6): 621–8.

Fiske, S. T., & Taylor, S. E. 2013. *Social Cognition: From Brains to Culture*. London: Sage Press.
Gendler, T. S. 2008. Alief and Belief. *Journal of Philosophy* 105: 634–63.
Gendler, T. 2011. On the epistemic costs of implicit bias. *Philosophical Studies* 156: 33–63.
Gino, F., & Galinsky, A. D. 2012. Vicarious dishonesty: When psychological closeness creates distance from one's moral compass. *Organizational Behavior and Human Decision Processes* 119: 15–26.
Graham, J., Haidt, J., & Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96: 1029–46.
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101: 366–85.
Greene, J. D. 2008. The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology*, 35–80. Cambridge, MA: MIT Press.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74: 1464–80.
Haidt, J. 2001. The emotional dog and its rational tail. *Psychological Review* 108: 814–34.
Haidt, J., Bjorklund, F. & Murphy, S. 2000. Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, University of Virginia.
Henrich, J., Heine, S. J., & Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33: 61–83.
Hilton, D. J. 1995. The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin* 118: 248–8.
Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. 2014. Morality in everyday life. *Science* 345: 1340–3.
Hutchings, V. L. 2009. Change or more of the same? Evaluating racial attitudes in the Obama era. *Public Opinion Quarterly* 73: 917–42.
Isserow, J., & Klein, C. 2017. Hypocrisy and Moral Authority. *Journal of Ethics and Social Philosophy* 12: 191–222.
Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Macmillan.
Kelman, M. 2013. Moral realism and the heuristics debate. *Journal of Legal Analysis* 5: 339–97.
Kennett, J. 2011. Imagining reasons. *Southern Journal of Philosophy* 49: 181–92.
Kihlstrom, J. F. 2004. Is there a "People are Stupid" school in social psychology? *Behavioral and Brain Sciences* 27: 348.
Kihlstrom, J. F. 2008. The automaticity juggernaut—or, are we automatons after all? In J. C. Kaufman, R. F. Baumeister, & J. Baer (Eds.), *Are We Free? Psychology and Free Will*, 155–80. New York: Oxford University Press.
Kohlberg, L. 1969. Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of Socialization Theory and Research*, 110–20. Chicago, IL: Rand McNally.
Laden, A. 2010. Reasoning: It's not all in the head. *Human Development* 53: 105–9.
Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347: 262–5.
Levy, N. 2014. Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs* 49: 800–23.
Liu, B. S., & Ditto, P. H. 2013. What Dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science* 4: 316–23.
Machery, E., Kelly, D., & Stich, S. P. 2005. Moral realism and cross-cultural normative diversity. *Behavioral and Brain Sciences* 28: 830.
Milgram, S. 1963. Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67(4): 371–8.

Norenzayan, A., & Schwarz, N. 1999. Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology* 29: 1011–20.

Orne, M. T. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17: 776–83.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J. & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105(2): 171.

Piaget, J. 1932. *The Moral Judgement of the Child*. London: Routledge & Kegan Paul.

Racial attitudes survey. 2012, October 29. The Associated Press. http://surveys.ap.org/data%5CGfK%5CAP_Racial_Attitudes_Topline_09182012.pdf

Royzman, E. B., Kim, K., Leeman, R. F. et al. 2015. The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making* 10: 296–313.

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. 1984. Pre- and postlexical loci of contextual effects on word recognition. *Memory & Cognition* 12: 315–28.

Simon, H. A. 1996. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Stanovich, K. E., & West, R. F. 1983. On priming by a sentence context. *Journal of Experimental Psychology: General* 112(1): 1–36. doi: https://doi.org/10.1037/0096-3445.112.1.1

Stein, E. 2005. Wide reflective equilibrium as an answer to an objection to moral heuristics. *Behavioral and Brain Sciences* 28: 561–2.

Sterelny, K. 2012. *The Evolved Apprentice*. Cambridge, MA: MIT Press.

Strohminger, N., Caldwell, B., Cameron, D., Borg, J. S., & Sinnott-Armstrong, W. 2014. Implicit morality: A methodological survey. In C. Luetge, H. Rusch, & M. Uhl (Eds.), *Experimental Ethics: Towards an Empirical Moral Philosophy*, 133–56. London: Palgrave Macmillan.

Sunstein, C. R. 2005. Moral heuristics. *Behavioral and Brain Sciences* 28: 531–41.

Tetlock, P. E. 1992. The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology* 25: 331–76.

Turiel, E. 2010. The development of morality: Reasoning, emotions, and resistance. In R. M. Lerner, M. E. Lamb, & A. M. Freund (Eds.), *The Handbook of Life-Span Development*, 554–83. New York: John Wiley and Sons.

Wallace, R. J. 2010. Hypocrisy, moral address, and the equal standing of persons. *Philosophy & Public Affairs* 38: 307–41.

Wielenberg, E. J. 2010. On the evolutionary debunking of morality. *Ethics* 120: 441–64.

Zulawski, D. E., & Wicklander, D. E. 2002. *Practical Aspects of Interview and Interrogation*. New York: CRC Press.

Further Readings for Part V

Demaree-Cotton, J., & Kahane, G. (2019). The neuroscience of moral judgment. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), *The Routledge Handbook of Moral Epistemology*, 84–149. New York: Routledge.

Argues that neuroscience is typically relevant to moral epistemology only when it bears on cognitive science claims pitched at higher levels of descriptions and evaluates extant attempts to draw conclusions for moral epistemology on the basis of theories in the cognitive science of morality.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, 2(3), 528–54. https://doi.org/10.1111/j.1756-8765.2010.01094.x

Attempts to derive conclusions about ethics from a theory on which moral behavior is shaped by social heuristics tailored to the social environment.

Kauppinen, A. (2014). Ethics and empirical psychology – critical remarks to empirically informed ethics. In M. Christen, C. van Schaik, J. Fischer, M. Huppenbauer, & C. Tanner (Eds.), *Empirically Informed Ethics: Morality between Facts and Norms*, 279–305. https://doi.org/10.1007/978-3-319-01369-5_16

Provides a pessimistic assessment of attempts to derive metaethical and ethical conclusions from empirical work.

Nichols, S. (2014). Process debunking and ethics. *Ethics*, 124(4), 727–49. https://doi.org/10.1086/675877

Provides an account of the conditions under which empirically based debunking arguments in metaethics and ethics are likely to be successful or unsuccessful.

Plakias, A. (2016). Metaethics: Traditional and empirical approaches. In J. Systsma & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy*. doi: 10.1002/9781118661666.ch13

Reviews attempts to use empirical work in moral psychology to make progress on debates in metaethics regarding moral realism, cognitivism, and moral motivation.

Pölzler, T. (2018). Can the empirical sciences contribute to the moral realism/anti-realism debate? *Synthese*, 195(11), 4907–30. https://doi.org/10.1007/s11229-017-1434-8

Provides an optimistic assessment of attempts to use empirical work to adjudicate debates in metaethics, including debates between cognitivism and non-cognitivism, and provides an account of the conditions that must be met in order for empirical work to do so.

Prinz, J. (2015). *An Empirical Case for Motivational Internalism*. Retrieved from https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199367955.001.0001/acprof-9780199367955-chapter-4

Provides several empirically based arguments in favor of motivational internalism about moral judgment.

Rini, R. A. (2015). Psychology and the aims of normative ethics. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics*, 149–68. https://doi.org/10.1007/978-94-007-4707-4_162

Defends the relevance of empirical work to adjudicating debates on normative ethics, the nature of moral agency, and the extent to which people are justified in holding their moral beliefs.

Tiberius, V. (2006). Well-being: Psychological research for philosophers. *Philosophy Compass*, 1(5), 493–505.

Discusses interactions between the philosophical and psychological study of wellbeing, while considering the relevance of this research for public policy.

Wright, J. C. (2018). The fact and function of meta-ethical pluralism: Exploring the evidence. *Oxford Studies in Experimental Philosophy*, 2, 119–50.

Argues that empirical work shows that most people are both realists and anti-realists about morality, and that this has certain benefits for society.

Study Questions for Part V

1) According to Kumar, why does empirical work on moral judgment favor his hybrid view of moral judgment over competing views? What assumptions does this argument rely on?
2) Kumar argues that empirical work can inform non-ideal ethical theory. What crucial normative assumption does this argument depend on?
3) What kind of moral reasoning does Kumar argue is appropriate for non-ideal ethics? What empirical work does he cite to support this view? Why would Klein's interpretation of this empirical work, if correct, cut against Kumar's argument for this kind of reasoning?
4) According to Kumar, what distinguishes debunking arguments that are likely to be successful from extant debunking arguments?
5) What is the difference between the Motivated Tactician view and the Activated Actor view?
6) How would a non-ideal ethical theory informed by the Motivated Tactician view differ from a non-ideal ethical theory informed by the Activated Actor view?

Index

abduction 39–41
Activated Actors 175–9, 183–4
Adams, F. 5, 108–15
adaptationism 91
affordances 91–3, 97–8, 100–2
algorithms 22, 91, 128, 137
amodal cognition 59, 110, 113
Anderson, E. 166, 168
Anderson, M. 4–5, 88–9, 94–6, 110–12
Anderson, S. 34
Appiah, K. 168
Aristotelian ethics 167
attention 138, 177–8
attitude 157

b-formatted representations *see* bodily formatted representations
Bakan, D. 183
Bargh, J. A. 177
Barsalou, L. 113
Bayesian abduction 39–40
Bayesian statistical learning 39–40
behavior 92, 159; brain-body-environment 52–7, 59; development 54–7; developmental pathways 57–8; perception-action separation 88–9, 100–2, 109
behavioral research 134–5, 139–41, 143–4; attention 138; contemporary disdain 138–9; experimental design 141–3; face perception 139–40; language 140; learning 136–7; memory 137–8, 140, 142–3; results concerning the brain 135–8; theory of mind 140; visual cognition 135
behaviorism 122
beliefs 157, 159–60
Berlinsky, D. 21
Bhalla, M. 100–1
biological adaptation 161
bodily formatted representations 88–9, 98–102, 109–10, 114
body: brain-body-environment networks 52–7, 59, 65–8, 76–8; development 54–7; *see also* embodied cognition
Bowlby, J. 91
brain *see* behavioral research; neuroscience
brain networks 50–2; brain-body-environment 52–7, 59, 65–8, 76–8; change 50–2, 58; cognitive system 76–7; development 54–7; developmental pathways 57–8; interaction 51; *see also* dynamic systems; embodied cognition
Brooks, D. 176
Brooks, R. 93
Brownstein, M. 165
Byrge, L. 49–59, 63–5, 70–2, 74, 76–8

Campbell, R. 164, 169
Caramazza, A. 113–14
Carey, S. 75–6
Casasanto, D. 94
causal buffering 67–70, 74–5, 78
causal mutualism *see* mutualism
causation 141–3
central cognitive process/reasoning system 77–8, 88–9
character traits 167–8
Chomsky, N. 14, 16, 18–19, 21, 25, 27, 29–30, 38
Cisek, P. 96–7
citation analysis 128–32
classical cognitivism 108, 112–15
co-instantiation 63–4, 71–5
Cognition 128–32
cognitive neuroscience *see* neuroscience
cognitive representation 108, 112, 124; *see also* representations
cognitive research 121–2, 127, 139–41; quantitative citation analysis 128–32; *see also* behavioral research

cognitive-brain-body-environment networks 76–7
cognitivism/non-cognitivism debate 156–60
computation 70–7
computational level 137
computational models 137–40, 143–4
computational psycholinguistics 38–9
computational theory of cognition 70–1, 137
concept empiricism 94–5
concepts 49, 59, 75–6
conceptual metaphor theory 95
connectivity 51
consistency reasoning 169, 178–82, 184
constitution 108–11, 113, 157–60
constitutive mutualism *see* mutualism
conversational norms 181–2
correlation 129, 136, 141–2, 183
Crain, S. 26, 33
Cronbach, L. J. 183
Culicover, P. W. 32
cultural adaptation 161
curiosity 40
Cushman, F. 6, 121–32, 138–9, 144

Dabrowska, E. 32
Danziger, S. 164
debunking arguments 161–5
dedicated input analyzers 75–8
degeneracy 72–5
deontology 160–3, 166
development: brain-body-behavior networks 54–7, 59; embodied cognition 90–3; innateness 68–70; moral 183–4
developmental niches 57
developmental pathways 49, 57–9
developmental psychology 33, 54
Dewey, J. 89, 91, 111
Dijkstra, K. 94
discrimination 165, 176–8
disgust 163–4
Ditto, P. H. 181
domain specificity 140
Doris, J. 167
Designer Receptors Exclusively Activated by Designer Drugs (DREADDs) 135
dualism 59
dynamic systems 77–8; causal buffering 67–70, 74–5, 78; co-instantiation 71–5; cognitive-brain-body-environment networks 76–7; computation 70–7; degeneracy 72–5; innateness 68–70, 75–6; metaphysical transducers 67–70, 73, 75–6; metaphysics 65–8; *see also* brain networks

Edelman, G. M. 72
EEG 122, 127
egocentric vision 55–7
eliminativism 157
embodied cognition 87–8; affordances 91–3, 97–8, 100–2; evolutionary hypothesis 90–3, 96; grounded cognition 94–6; neural reuse 94–8; strong account 88–90, 110–12, 114–15; Transiently Assembled Local Neural Subsystems (TALoNS) 96–8, 111–12; weak account 87–9, 98–102, 108–10, 112–15; *see also* neural reuse
emotionism 158–9
empirical induction 22, 24, 37
environment 90–2; affordances 91–3, 97–8, 100–2; brain-body-environment networks 52–7, 59, 65–8, 76–8; embodied cognition 90–3, 96–8, 111
Environment of Evolutionary Adaptedness (EEA) 90–1
epistemology 40, 160–1
ethics 155–6, 170; consistency reasoning 169, 178–82, 184; debunking arguments 161–5; deontology 160–3, 166; disgust 163–4; epistemology 160–1; evolutionary psychology 161–2; moral development 183–4; moral dumbfounding 159, 180–3; moral intuitions 162–4; moral judgment 156–60; moral knowledge 158; moral reasoning 168–9, 174–84; non-ideal theory 165–70; utilitarianism 160–1, 163, 166
Evans, N. 32
evolution 90–3
evolutionary psychology 161–2
experience 14–15, 18–21, 24–6, 31, 51; *see also* brain networks
experimental situations 181–3
explanation 73, 122–3, 130–1, 159–60, 166–7, 174–8, 181–3
extended cognition 108

face perception 139–40
Faculty of Language (FL) 13–14
Fedyk, M. 3–4, 63–78
Feigl, H. 40
Fiebelkorn, I. C. 138
Fiske, S. T. 175–8, 183–4
Fodor, J. A. 59, 63
framing effects 162
Friston, K. J. 40

functional location 63
functional networks 50–1; *see also* brain networks

Gallagher, S. 102
Gallese, V. 95, 99
generalizations 22–5, 34–6
Gibson, J. J. 91
Glenberg, A. M. 94, 111, 113
Goldman, A. 4–5, 87–90, 98–102, 108–15
Gollwitzer, P. 168
Golonka, S. 90
Goodman, N. 22–4
Goodwin, G. 163–4
grammars 13–14; *see also* Universal Grammar
Greene, J. 162–3
Gricean norms 181–2
grounded cognition 95–6
grue 22–5, 37

Haidt, J. 180–2
Harris, Z. 38–9
heuristics and biases 162–4, 176, 178, 181
Higginbotham, J. 20
Hodgkin–Huxley model 139
homophony 17–20, 24–6
Hornstein, N. 1–3, 13–27, 29–41
Hubel, D. H. 125, 127–8
Hurley, S. 99, 108, 112
hybrid theory (metaethics) 159–60
hypocrisy 179–80

image schemas 95
individual differences 183
induction 22–5, 27, 37–9
information processing 77
Ingold, T. 92, 111
innateness 49, 59, 75–8, 176; capacities 27; causal buffering 67–70, 74–5, 78; concepts 70, 75–6; degeneracy 72–5, 78; as developmental essentiality 68–70; *see also* Universal Grammar
inner mental representations 29–30, 51, 87–91, 122, 176–7; *see also* representations
interference and competition effects 94–5
Isserow, J. 179
is–ought problem 155, 161, 170

James, W. 51–2, 58
judgment *see* moral judgment

Kalaska, J. F. 96–7
Kamin, L. 136–7
Kaschak, M. P. 94, 111
Kelly, D. 163
Kennett, J. 181
King, H. V. 34
Kirby, S. 41
Kiverstein, J. 4–5, 87–102, 108–15
Klein, C. 8, 174–84
Knobe, J. 159
knowledge 40, 160–1
Kohlberg, L. 183
Kumar, V. 7–8, 155–70, 174–5, 178, 180, 184

Lakoff, G. 95
Landy, J. 163–4
language 15, 98, 140
language *see* Universal Grammar
language acquisition 32–3, 41
Language Acquisition Device 15–16, 19, 21
Lasnik, H. 16
learning *see* brain networks
learning: algorithms 22, 91, 128, 137; Bayesian statistical learning 39–40; prediction error 137; reinforcement learning 49, 137
lesion 122, 136, 139–42
lesion studies 140–1
Leslie, S.-J. 177
levels of analysis 73
Levinson, S. 32
Lewontin, R. 91
lexical meaning 33, 37
Lightfoot, D. W. 34
limited homophony 19
linguistic creativity 13–14
linguistic diversity 30–2
linguistic nativism 32
linguistics *see* Universal Grammar
literacy 57
Liu, B. S. 181
logic 22–4, 29–30, 77

McGaugh, J. L. 136
magnetic resonance imaging (MRI) 122, 127
Mahon, B. Z. 113–14
Marr, D. 76–7, 137
massive modularity hypothesis (MMH) 90–1, 93
May, J. 162–4
memory recall 94, 137–8, 140
mental representation *see* inner mental representation
metaethics 155–61, 165; cognitivism/non-cognitivism debate 156–60; hybrid theory 159–60

metaphysical transducers 67–70, 73, 75–6
metaphysics of mechanisms 66
methodology: causal 108–13, 136, 141–2, 158–60, 165; correlational 136, 141–2; empirical 22–5, 69–72, 75–7, 113, 155, 157–65, 167–70, 181–2; a priori 74, 123–5, 143–4, 156, 165
Milgram, S. 181
mind *see* cognitive research; dynamic systems
modal cognition 112, 142–3
modules 90–1, 102–3, 109, 139
Molaison, H. 140
moral consistency reasoning 169, 178–82, 184
moral development 183–4
moral dumbfounding 159, 180–3
moral intuitions 162–4
moral judgment 156–60
moral knowledge 158
moral reasoning 168–9, 174–5; consistency reasoning 169, 178–82, 184; Motivated Tacticians 175–8, 180–1, 183–4
Morsella, E. 177
Motivated Tacticians 175–8, 180–1, 183–4
motivations 179–80
mutualism 91, 111

natural psychological kinds 157–8, 160
naturalistic ethics 157–61, 170
neural function 125, 139
neural implementation 142–3
neural reuse 90–4; embodied cognition 94–8; moderate account (Goldman) 98–102; simulation-based account 99–100; *see also* embodied cognition
neural structure 72, 94, 98, 135
neuroscience 121–3, 132, 138–41; behavioral research 134–6, 138–44; causal manipulations bias 141–3; importance 123; lesion studies 140–1; a priori analysis 123–5; quantitative citation analysis 128–32; relevance 124; theory and methods distinction 124–5, 127–8; visual cognition case study 125–8
neuroscientific methods: EEG 122, 127; lesion 122, 136, 139–42; MRI 122, 127; PET 127; single-unit recording 122, 125, 127; transcranial stimulation 141
neuroscientifically relevant psychological (NRP) factors 96–8
Niv, Y. 6–7, 134–44
Nola, R. 39
non-ideal theory 165–70
Norenzayan, A. 182
normative ethics 155–6, 161, 163, 165–70, 181

object manipulation 54, 58
ontology 73–4
optogenetics 135
Orne, M. T. 181

Packard, M. G. 136
pathways *see* developmental pathways
perception-action separation 88–9, 100–2, 109
permissive grammar 19–20, 26
Pessoa, L. 94
PET 127
phenotypic plasticity 92
Piaget, J. 54, 183
Pietroski, P. 1–3, 13–27, 29–41
power 177–8
pragmatics 36, 181
prediction error 137
primary linguistic data (PLD) 14, 21
priming 177
Prinz, J. 158–60
probabilistic reasoning 39–40, 77
procedures 15–21, 25
Proffitt, D. 100–1
projection 20–1, 23–5, 27
pronunciation-meaning pairs (p-μ pairs) 15–17, 20, 27
psychological computation 77; *see also* computational models
psychophysics 50, 135, 138
Pullum, G. 2–3, 29–41
Pulvermüller, F. 97–8, 112

racism 174–6
Railton, P. 167–8
rat maze study 136
rationality 77–8, 175
Rawls, J. 166, 168
reductionism 59, 134
reinforcement learning 49, 137
rejoinder emphasis 34–5
representations: bodily formatted 88–9, 98–102, 109–10, 114; cognitive 108, 112, 124; inner mental 29–30, 51, 87–91, 122, 176–7
Rescorla, R. 137
restrictive grammar 18–20
retinal ganglion cells 125–7
retrieval-induced forgetting 137–8
Royzman, E. B. 181, 183
Rupert, R. 102

sampling 25, 54, 57, 138
sandwich view of cognition 108, 112–15
Sankey, H. 39
Schwarz, N. 182
self-control 168
semantics 37–9
Sen, A. 166
sensory-motor skills 55–7; *see also* embodied cognition
Shapiro, L. 110, 112, 114–15
single-unit recording 122, 125, 127
Sinigaglia, C. 99
Sinnott-Armstrong, W. 162, 164
Smith, L. 3, 49–59, 63–5, 70–2, 74, 76–8
social psychology 174, 184; Activated Actors 175–9, 183–4; experimental situations 181–3; local consistency reasoning 180–4; moral development 183–4; Motivated Tacticians 175–8, 180–1, 183–4
Sporns, O. 49–59, 63–5, 70–2, 74, 76–8
stereotypes 177–8
stimulation (magnetic/electrical) 122, 139, 141
structural networks 50–1; *see also* brain networks
sui generis cognitive system 77
Symons, D. 91
syntax 31–4, 37–9
systems *see* dynamic systems

T-shaped maze study 136
Taylor, S. E. 175, 184
Tetlock, P. E. 182
theory of mind 140
traditional view of cognition 108, 112–15
transcranial electrical stimulation 122, 139, 141
transcranial magnetic stimulation 122, 139, 141
Transiently Assembled Local Neural Subsystems (TALoNS) 96–8
Turiel, E. 183–4
Turing, A. 70–2

uncertainty 40
unconscious influences/processing 155, 162–3, 165, 168, 174–5, 178
universal availability 13–14
Universal Grammar 13–15, 25–7; abduction 39–41; alternative language acquisition literature 32–3, 41; contextual plausibility 35–6; developmental psychology 33; experience 14–15, 18–21, 24–6, 31, 36–7; generalizations 22–5, 34–6; homophony 17–20, 24–6; induction 22–5, 27, 37–9; lexical meaning 33, 37; linguistic diversity 30–2; logic 29–30; procedures 15–21, 25; projection 20–1, 23–5, 27; pronunciation-meaning pairs (p-μ pairs) 15–17, 20, 27; universals 19–21
universals 19–21
utilitarianism 160–1, 163, 166

virtue ethics 167–8
visual cognition 125–8, 135

Wagner, A. 137
Wallace, R. J. 179
Walsh, D. 92
West-Eberhard, M. J. 92
Wicklander, D. E. 179
wide reflective equilibrium 168–9
Wiesel, T. N. 125, 127–8
Wilson, A. 90
working memory 138, 142–3

Xu, F. 3–4, 63–78

Zulawski, D. E. 179
Zwicky, A. M. 34



Taylor & Francis eBooks

www.taylorfrancis.com

A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.

90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.

TAYLOR & FRANCIS EBOOKS OFFERS:

- A streamlined experience for our library customers
- A single point of discovery for all of our eBook content
- Improved search and discovery of content at both book and chapter level

REQUEST A FREE TRIAL

support@taylorfrancis.com

Routledge
Taylor & Francis Group

CRC Press
Taylor & Francis Group