# THE SHAPE OF THOUGHT

## How Mental Adaptations Evolve

H. CLARK BARRETT

# The Shape of Thought

Evolution and Cognition
General Editor: Stephen Stich, Rutgers University

Simple Heuristics That Make Us Smart
Gerd Gigerenzer, Peter M. Todd, and ABC Research Group

Adaptive Thinking
Rationality in the Real World
Gerd Gigerenzer

Natural Selection and Social Theory
Selected Papers of Robert Trivers
Robert Trivers

In Gods We Trust
The Evolutionary Landscape of Religion
Scott Atran

The Origin and Evolution of Cultures
Robert Boyd and Peter J. Richerson

The Innate Mind
Volume 2: Culture and Cognition
Edited by Peter Carruthers, Stephen Laurence, and Stephen Stich

Why Humans Cooperate
A Cultural and Evolutionary Explanation
Joseph Henrich and Natalie Henrich

The Innate Mind
Volume 3: Foundations and the Future
Edited by Peter Carruthers, Stephen Laurence, and Stephen Stich

Rationality for Mortals
How People Cope with Uncertainty
Gerd Gigerenzer

Simple Heuristics in a Social World
Ralph Hertwig, Ulrich Hoffrage, and ABC Research Group

The Shape of Thought
How Mental Adaptations Evolve
H. Clark Barrett

# The Shape of Thought

## *How Mental Adaptations Evolve*

H. Clark Barrett

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.






Library of Congress Cataloging-in-Publication Data
Barrett, H. Clark.
The shape of thought : how mental adaptations evolve / H. Clark Barrett.
pages cm.—(Evolution and cognition series)
Includes bibliographical references and index.
ISBN 978–0–19–934831–2 (pbk. : alk. paper)—ISBN 978–0–19–934830–5 (hardcover : alk. paper)
1. Thought and thinking. I. Title.
BF441.B2963 2014
153.4′2—dc23
2014007239

The manufacturer's authorized representative in the EU for product safety is
Oxford University Press España S.A. of Parque Empresarial San Fernando de Henares,
Avenida de Castilla, 2 – 28830 Madrid (www.oup.es/en or product.safety@oup.com).
OUP España S.A. also acts as importer into Spain of products made by the manufacturer.

Printed in the United States of America on acid-free paper

# CONTENTS

Acknowledgments vii

**Introduction: The Problem 1**

**Part I: Evolution**
**1. Additivism 17**
**2. Hill-Climbing 50**

**Part II: Information**
**3. Adding Value 85**
**4. Social Ontology 104**
**5. Minds 128**

**Part III: Development**
**6. Development 155**
**7. Open Ends 182**

**Part IV: Culture**
**8. Moving Targets 209**
**9. Culture 224**
**10. Accumulation 243**

**Part V: Architecture**
**11. Parts 261**
**12. Wholes 289**
**13. Us 320**

**Conclusion: Possibilities 333**

References 339
Index 379

# ACKNOWLEDGMENTS

Thanks to all those whose ideas, conversations, comments, and support made this a better book, sometimes in ways unbeknownst to them. There are far too many to name, but I'd especially like to thank Renée Baillargeon, Nikki Bazar, Rob Boyd, Pascal Boyer, Gary Brase and his evolution and cognition seminar, Greg Bryant, Peter Carruthers, Leda Cosmides, Gergo Csibra, Mirella Dapretto, Tom Dickins, Dan Fessler, Alan Fiske, Tom Flamson, Willem Frankenhuis, Gyuri Gergely, Tamsin German, Matt Gervais, Ray Gibbs, Gerd Gigerenzer, Martie Haselton, Joe Henrich, Michelle Kline, Rob Kurzban, Steve Laurence, Deb Lieberman, Philip Lord, Joe Manson, Sarah Mathew, Cristina Moya, Siamak Naficy, Karthik Panchanathan, Susan Perry, Elizabeth Pillsworth, Annemie Ploeger, Roberto Schonmann, Brooke Scelza, Joan Silk, Jeff Snyder, Steve Stich, Dan Sperber, Don Symons, Lisa Thorne, Jason Throop, Peter Todd, John Tooby, Andreas Wilke, participants in the fall 2013 evolutionary psychology seminar at Central European University, and all my colleagues and students at UCLA. Thanks to James Turrell for permission to use his work *Afrum* (White), 1966, on the cover (Photograph by Florian Holzherr). And finally, I'd like to thank my family: Mom, Dad, Joyce, Chris, and Katie.

# INTRODUCTION
## THE PROBLEM

There are few things more obvious than the fact that we are the products of evolution. Like everything else alive on this planet, we are the result of descent with modification over evolutionary time. This goes for our bodies, of course—our hands, our livers, our hearts, our skin—but it's also true of our minds. Of *course* they evolved; what else could explain them? How else could we hope to understand them except in the same way that we understand everything else on earth that is the product of evolution?

This much, it seems, is something with which virtually every modern-day scientist would agree. And yet, while scientists all agree *that* evolution shaped the mind, those same scientists can't seem to agree *how* it did. That is to say, while scientists agree that the mind (and the brain, which I will use largely synonymously with mind in this book) is a kind of machine made of neurons that takes in information, causes us to think, and makes our bodies move around, they can't agree on just how evolution has shaped what's inside the machine that causes it do those things.

You might be surprised to hear this, but it is true. In the present-day sciences of the mind—including psychology, anthropology, neuroscience, and other sciences of human behavior—there are few things a scientist can do to generate more controversy than to make a claim that this or that feature of human thought or behavior is the product of evolution. Nowhere is this more clear than in the barrage of criticism that continues to be directed at the field of evolutionary psychology—my own field of study and the subject of this book—which attempts to ask just how evolution has shaped what's inside the mind. As research in evolutionary psychology has increased in stature and visibility over the past two decades, so has the controversy and criticism swirling around it. So great is the skepticism regarding evolutionary psychology that one critic, philosopher David Buller, has

suggested that the most widely recognized work in the field is "wrong in almost every detail" (Buller, 2005, p. 3).[1]

If you are not a social scientist yourself, and if you agree that it's obvious that humans evolved, then you might find it puzzling that there should be controversy surrounding the idea of bringing evolutionary theory to the study of the mind. Indeed, *that* much isn't really controversial at all. Most social scientists would probably agree with the statement that there should be "an" evolutionary psychology of some kind. But the evolutionary psychology that I will be endorsing in this book seeks to go beyond the mere truism that the mind evolved. It wants the details. It seeks to ask how, why, and in what way evolutionary processes, including natural selection, have created and shaped the mind's functional features. Such features—properties of organisms that do functional things in the service of survival and reproduction—are what biologists call "adaptations." And it turns out that when you start using that word to refer to things in the mind, things start to get ugly. Say that the mind evolved and you're fine; but point to some feature of the mind and call it an "adaptation"—even as a hypothesis to be tested—and you start to get in trouble.

Notoriously, the most extreme critics argue that adaptationist hypotheses are simply "just-so stories" that can rarely, if ever, be truly tested. As biologists Stephen Jay Gould and Richard Lewontin put it: "Since the range of adaptive stories is as wide as our minds are fertile, new stories can always be postulated" (Gould & Lewontin, 1979, p. 153). To some, this means that the business of proposing and testing adaptationist hypotheses (or at least pretending to) is a waste of time. If scientific progress is your thing, you might be a bit depressed to hear this. How can we distinguish between different possibilities for how and why the mind is organized as it is without considering what those possibilities are, and trying to adjudicate between them with a combination of theory and empirical evidence?

This is not to say that all proposals of adaptations in the mind are equally controversial. For example, it's relatively uncontroversial that perceptual mechanisms that convert external information into a format the brain can use, such as the cells in our auditory cortex that respond to different frequencies of sound—"sensory transducers"—are adaptations. And it's uncontroversial that basic emotions like fear and hunger and the brain structures responsible for them are the products of evolution, as well as systems of motor control and some other so-called peripheral or lower-level systems. But as soon as you start to move away from things like perception and motor

[1] For a small sample of the growing pile-on of criticism directed toward evolutionary psychology, see Bolhuis et al. (2011); Buller (2005); Gray et al. (2003); Dupré (2001); Fodor (2000); Lickliter & Honeycutt (2003); McCaughey (2007); McKinnon (2005); Panksepp & Panksepp (2000); Quartz & Sejnowski (1997); Richardson (2007); Rose & Rose (2000); Scher & Rauscher (2002).

control into so-called higher levels of cognition—what is colloquially called "thinking"—then the trouble really starts. In the realm of thinking, including phenomena like reasoning, judgment, decision-making, language—in short, virtually everything we do with information after it enters the mind—proposals about adaptations are massively controversial. This is a problem, because there seems no doubt that human thinking must be the product of natural selection too.

Evolutionary psychologists typically claim, as will I, that human thinking is likely to be carried out by many specialized adaptations, sometimes (though also controversially) called "modules" (Barrett & Kurzban, 2006; Pinker, 1997; Tooby & Cosmides, 1992). This position (often called "massive modularity"; Carruthers, 2006; Gibbs & Van Orden, 2010; Machery, 2007, 2008; Samuels, 1998; Sperber, 1994, 2001) is sometimes depicted as bordering on crazy. The originator of modularity theory himself, philosopher Jerry Fodor, has called it "modularity theory gone mad" (Fodor, 1987, p. 27). But if you look at the idea soberly—thought must be carried out by multiple, specialized adaptations—there is a logic behind it that makes it quite hard to avoid. *That* adaptations are involved follows from the idea that human abilities of thought are the products of natural selection. That there must be *multiple* adaptations follows from the observation, both empirically and theoretically based, that thought is not just a single thing produced by a single mechanism. Instead, much like other complex biological processes such as cell replication, metabolism, or the immune response, what might seem like a seamless phenomenon—thinking—is likely to be produced by an underlying battery of interacting mechanisms, each shaped to carry out a particular function. If that's true, the question becomes, what *are* those mechanisms? What is the evolutionary history of each, and what function has it been selected to carry out?

Yet many experts on the brain, including neuroscientists, claim that the parts of the brain largely responsible for "thinking"—in particular, the outer regions of the brain collectively known as the cortex—are most definitely *not* composed of adaptations. As neuroscientists Steven Quartz and Terrence Sejnowski put it, "the developing cerebral cortex is largely free of domain-specific structure" (Quartz & Sejnowski, 1997, p. 537). In other words, it contains no specialized adaptations. On this view, what it contains instead is *plasticity*—the ability to respond to circumstance in an adaptive way. Philosophers David Buller and Valerie Hardcastle put this position as follows: ". . . with the possible exception of our sensory transducers, it is not the contingently stable brain structures in an adult's brain that are adaptations; rather, the brain's very developmental plasticity is the adaptation" (2000, p. 321).[2]

[2] It's possible to find many similar statements in the literature. For example, neuroscientists Jaak and Jules Panksepp state: "The organization of the neocortex, although still constrained by many unknown genetic rules, may be much more of a general purpose computational

Let's be clear on the scope of disagreement here. Some supposed experts on the evolution of mind, such as myself and many other evolutionary psychologists, claim that the entire brain, including the cerebral cortex, is likely to be composed of adaptations. Others claim that essentially none of it is.

I hope you'll agree with me that this is quite a startling difference of opinion. Imagine a disagreement of this magnitude in some other field—say, chemistry or physics—over an issue so basic in that particular field of study. A: "Matter is composed of atoms." B: "No it's not." Good luck getting your car started in the morning, much less building a rocket that can land on the moon.[3]

So what has gone wrong? How can there possibly be so much difference of opinion among experts over such basic issues in their area of expertise?

The main reason, I'll claim, is misunderstanding: misunderstanding of just how evolution might shape the mind, and of what mental adaptations might be. I think the core of the problem lies in a failure to appreciate just what "specialization" means. From a biological point of view, the word "specialization" implies *diversity.* Think of a coral reef or a rainforest: Adaptations come in all shapes and sizes, each elegantly crafted but often bizarrely different. This is true both across organisms and within them. Insects, for example, have radiated out via a process of evolutionary diversification into countless shapes and lifestyles, from beetles to butterflies to bumblebees. And within organisms, diversity of specialization occurs as well, driven by the principle of division of labor. Limbs, lungs, and livers are not the same because they do different things in the service of survival and reproduction. And this is true at many levels. Dig down deep into any organism and you will see specialization nested within specialization. The "macro" function of the kidney is regulation of the blood, but it contains many adaptations nested within it that allow it to carry out even more specialized "micro" functions like regulating electrolytes, blood acidity, and blood pressure. The same nesting principle occurs even within cells, down to the organelles and the specialized biomolecules within them. There is every reason to think this is also true of the function we call "thought."

A basic principle that runs through all of this is that different specializations are, well, different. Locomotion and digestion, for example, are different problems,

device than modern evolutionary psychologists have been willing to concede . . . . The possibility is remote . . . that many unique and detailed epistemological engravings of sociobiological strategies (i.e., modules) exist within the human neocortex" (Panksepp & Panksepp, 2000, p. 110). For similar views, see Elman et al. (1996), Karmiloff-Smith (1992), Lickliter & Honeycutt (2003), Smith & Thelen (2003).

[3] This was in fact a debate that occurred in physics and chemistry not so long ago, at the beginning of the twentieth century. Look where we've come since then.

and nobody would expect the adaptations that solve them to share a common set of features. Nobody would expect adaptations in the heart to look like adaptations in the lungs, nor would they expect distinct adaptations within a given organ, like the light-focusing lens of the cornea and the light-sensitive rod and cone cells of the retina, to have the same design. It would be absurd to imagine that adaptations come in a single, cookie-cutter format; no biologist would propose some kind of prototype or template that all adaptations must follow. That flies in the very face of the idea of specialization.

And yet that cookie-cutter view is just how many psychologists think about specialized adaptations in the brain. They think that calling something a specialized adaptation implies that it carries with it a particular checklist of properties: in particular, it must be "innate," developing without learning or any other input from the environment or experience; it must be "domain-narrow," narrowly targeted toward some particular category of information; and it must be "automatic," or reflex-like, operating autonomously from other systems in the brain and independently of higher-level processes like reasoning and conscious choice, processes that we've called "thinking." These are properties typically attributed to what psychologists call "modules," which are widely seen to be synonymous with the idea of brain specializations, or adaptations (Fodor, 1983). Many psychologists contrast these kinds of mechanisms, which they believe to be the unconscious, modular, evolved reflexes of the mind, with the flexible, domain-general operations of consciousness and higher-level thought. This is sometimes called a dual-systems view (Evans, 2003; Kahneman, 2011; Stanovich, 2004).

Why would people think that adaptations come in only one cookie-cutter kind: innate, domain-narrow, isolated, inflexible, automatic modules? One reason is historical: In contemporary psychology, philosopher Jerry Fodor's definition of a module as akin to an innate cognitive reflex has been particularly influential (Fodor, 1983; see Barrett and Kurzban, 2006, for discussion). But the distinction between "instinct" and "reason" is ancient and can be traced from the debates of Plato and Aristotle, through Enlightenment philosophers like Descartes and Locke, to its present-day incarnation in the dual-systems distinction between "System 1" (modules, instincts) and "System 2" (domain-general mechanisms, reason). And aside from its historical roots, the idea is arguably a beguiling one due to its simplicity and intuitive appeal. In fact, it's relatively easy to make up an appealing story for why brain adaptations should be like this: innate, domain-narrow, and automatic. You can do this—although I'm arguing you shouldn't—by imagining a "prototypical" psychological adaptation, and then using it as a model for psychological adaptations in general.

Take fear of snakes. Snakes are a paradigm case of a fitness threat that was present in ancestral environments. Therefore, natural selection could have engineered specialized adaptations to deal with them. As it happens, snakes share certain perceptual features (e.g., they are squiggly and move in slithery ways). This means that natural selection could have engineered an innately specified "template" (almost like a little picture of a snake in the brain) that could be used to detect snakes. This could be wired up to a reflex-like reaction pathway that quickly and automatically makes you jump back upon seeing a snake. These features all make adaptive sense: Innateness of the response means you don't have to learn that snakes are dangerous, and being automatic, fast, and reflex-like—operating outside of conscious choice—means you don't waste time pondering your different options. And, as it happens, there is evidence for an evolved snake response with just these properties (LeDoux, 2000; LoBue et al, 2010; Öhman, 2005).

If you thought that this is what it means to call something a specialized adaptation—that it is innate, domain-narrow, and reflex-like—then claims by people like Sejnowski, Quartz, Buller, and Hardcastle begin to make sense. In fact, you don't even need to be a brain scientist to realize that there are many aspects of human thought that such a model of specialized adaptations can't account for. Much human thinking, for example, involves learning, rather than strict innateness. It involves interaction between systems, rather than isolation between them. It involves the ability to handle diverse classes of information, rather than narrowly targeted domains. It involves the ability to handle evolutionarily novel information that never existed in the past. And certainly, much of human thought involves processes like reasoning and choice. Given that these are all *opposites* of properties that specialized adaptations are supposed to have, then—so goes this line of thinking—they must be carried out by something other than specialized adaptations.

In present-day psychology, these other "general-purpose" or "domain-general" mechanisms are whatever specialized mechanisms aren't. They are flexible, able to deal with many kinds of information, and not innately specified (or at least not rigidly so; they adjust via experience). And it is widely believed that *most* of human cognition—including higher-order processes like reasoning, judgment, and decision-making—must be carried out not by adaptive specializations, but by these general-purpose mechanisms. Dual-systems theory is just one formalization of this view, commonly held in psychology, of a mind composed of two kinds of things: the specialized stuff and the general-purpose stuff.

Unfortunately, I think, it's likely to be wrong. Wrong not in suggesting that the mechanisms that underlie processes like reasoning and decision-making are *different* than those underlying perception and motor control; that's almost certainly true. What's wrong is the idea that these mechanisms are *not also specialized adaptations*. In other

words, it's not that some of the mind's mechanisms are specialized and others aren't, as the dualist view would have it; it's that they are all specialized to do different things.[4]

If this is true, then we should expect the diverse adaptations of the mind to have diverse properties, rather than coming in just two flavors. What it means to be "specialized" for the task of learning, for example, is not the same as what it means to be "specialized" for the tasks of vision or motor control. Properties like speed and automaticity might make sense for detecting snakes, but not necessarily for processes that happen slowly and use lots of information, such as learning a language. And some processes, like decision-making, might be *expected* to use information from many sources, rather than a single narrow domain. But just because the process isn't fast, automatic, or domain-narrow doesn't mean that adaptations aren't involved. No matter what the process—learning, reasoning, decision-making—if our brains have been selected to do it, that implies adaptations that get it done. As scientists, our job is to look for those adaptations and to find out what properties they might have, rather than deciding beforehand. Taking the template of a Fodorian "module" and assuming that it is what all brain adaptations must look like makes as much sense as looking for wings in our lungs or light-detecting cells in our kidneys.

The view I would like to explore, then, is quite different from the one that partitions the mind into two kinds of stuff. It is a view that expects the entire mind to be composed of adaptations, but that also expects those adaptations to be diverse. On this view, there is no line that we can draw across mental processes such that some portion of them, X, is handled by adaptations, and some other portion, Y, isn't. Instead, our entryway into the study of the mind should be the expectation—perhaps wrong in some cases, but a good starting point—that *all* aspects of thought that exhibit functional organization are likely to be the result of natural selection at some level, just as is the case for functionally organized stuff elsewhere in biology. This holds true from lower-level processes like perception, all the way up to the highest levels of thought and even consciousness itself. The relevant question is not *whether* adaptations are involved, but what those adaptations are, and what properties they have that allow them to do what they do. There are many ways in which the mind is "fitted," or shaped, to the world in which it operates, and all aspects of mind-world fit must, ultimately, be the result of the evolutionary process. Our goal is to figure out

[4] Most dual-systems theorists would probably agree that "System 2" mechanisms are the products of evolution, and possibly adaptations. However, it's common to call them "general-purpose," as if they have no particular functions or design features evolved to carry out those functions. They would rarely, if ever, be called "modules," despite the fact that many models of higher-level cognitive processes, like working memory, are "modular" in the sense of being composed of specialized components (e.g., Baddeley, 2003).

the details of how mind-world fit evolves, how it manifests in the brain, and how it is built anew each generation through the process of development.

In order to carry out such a research program, we are going to have to be true to the idea of diversity. We are going to have to be prepared for a mind that is not neatly orderable into two categories, but is rather a mish-mash. Among other things, the human brain is almost certainly a *mosaic* of evolved adaptations, some very old in evolutionary terms and some quite new. The story of how each element of this mosaic evolved is not likely to be simple. Because evolution proceeds by descent with modification, it is likely to involve many twists and turns. Some of these may be hard or even impossible to reconstruct. But if we are to have a hope of doing so, we must be open to all the possible paths that evolution might have taken. We shouldn't necessarily expect the "optimal" answer, nor should we expect the simplest or most parsimonious one, because the pathways that evolution takes are rarely the simplest or most obvious ones; indeed, some biologists would say they almost never are.

Instead, I'd like to suggest that there is really only one principle that holds true of all adaptations—a principle so general that I am going to call it a "law." It's an underappreciated fact about biology that there are actually precious few things that merit being called laws in the sense that we might speak of, say, the laws of physics. The reason is that evolutionary processes are inherently messy. They are historical, involving change over time, so prior states of the system are a necessary part of the explanation of present states. They occur over populations rather than single individuals, and populations are more like clouds than points: variable and spread out over time and space. And evolutionary processes are stochastic, or probabilistic. Evolution writ large is actually the aggregate of lots of things happening to lots of individuals over lots of time, and every specific one of those happenings is unique. For example, while there are general tradeoffs that shape the evolutionary process, such as quality-quantity tradeoffs (e.g., it's not possible to produce an infinite number of offspring with infinite life spans), this doesn't mean that the tradeoff is executed with mathematical precision in every case. Instead, it's only in the aggregate that anything remotely law-like takes place. Thus, while it is true that processes such as adaptation occur over time, and therefore things like form-function fit evolve, the details of any given case depend on—well, the details of that case.

So, not without a certain sense of irony, I'd like to give this idea a name: the first law of adaptationism. This law can be summarized as follows: *It depends.* In its long form, what this means is that the nature of any given adaptation—the form that it takes and the nature of its fit to the world—depends on how and why it got there. In fancier and more technical terms, the properties of an adaptation will depend on its functions, on the historical trajectory of descent with modification that gave rise to it, the tradeoffs involved in shaping its particular fit to its circumstances, and

the sources of stochasticity that generated the variation that selection acted upon. In short, the first law is that there are no laws—except itself!—that hold, rigidly and in the same way, across all adaptations. Instead, to properly understand the nature of any particular case in which there is a fit between the mind and the world, we will have to look carefully at the details of the case.

If we want to have a hope of understanding the entire mind in evolutionary terms, we're going to have to take this law seriously. The first law matters, and it matters a lot. It matters because if you prejudge the issue of what counts as an adaptation, or decide beforehand what you expect adaptations in the mind to look like, then you will never be able to capture all the ways that evolution has shaped the mind.

Let me give you a few examples. Consider the phenomenon of consciousness—one of the most contentious topics in psychology, but clearly an important part of our mental life. Remember that part of Fodor's definition of a module includes that modules must operate automatically, outside of conscious awareness. In snake detection, for example, there appears to be a kind of reflex loop that causes us to recoil in fear before any conscious choice. If you think that this is a property that must be true of adaptations *in general*, then by definition—by definition!—everything about conscious thought is outside the purview of adaptive explanation. This doesn't seem like a conclusion we'd like to reach. It seems unlikely to be true because consciousness itself, and the machinery that gives rise to our conscious choices, must also be the product of evolution by natural selection. Choices, even conscious ones, can affect fitness, and they always have. For these reasons, deciding beforehand that only lower-level processes like emotions and perception consist of adaptations leaves you unable to explain everything else in evolutionary terms.

A similar problem ensues if you equate "innate" with "evolved," such that only the features of our minds that develop invariantly or are present at birth are seen as the products of natural selection. This implies that features of our minds that vary across individuals, or that are learned, or that depend on experience in any way, must be the products of something else. Indeed, some would argue just this: As we saw above, Buller and Hardcastle have claimed that any aspect of the mind that is the product of experience cannot be an adaptation. But if that's true, then most if not all of our behavioral abilities—not just in humans but in other animals as well—cannot be adaptations, because they are shaped by experience. For example, this view would hold that the human ability to use language cannot be an adaptation, since it involves learning. But by the same token, neither can a leopard's ability to catch prey or a bird's ability to find her way back to her nest, since learning is involved in shaping those skills too.

As was the case with consciousness, I don't think this is a conclusion we want to draw. Clearly, our learning abilities themselves must be the products of natural

selection. But that's not all: These learning abilities have only evolved because of what they enable us to learn. That is why they exist. Therefore, their outcomes are not somehow separate from the adaptation. Just as bodily growth is a developmental process that builds wings and legs, learning is a developmental process that builds the human ability to speak and a bird's ability to find its way back to its nest; these are all parts of the organism's phenotype (the biological term that refers to the sum total of an organism's traits, the outcomes of development). And just as we wouldn't want to necessarily say that wings and legs aren't adaptations because there are developmental processes that build them, neither would we want to argue that just because a developmental process builds speech or navigation abilities (even if it uses input from the environment to do so) that those things cannot be adaptations. Moreover, if the function of learning mechanisms is to learn, then they must be specialized to learn and must have design features that allow them to do so. This means that the important adaptationist questions about learning mechanisms are: What have they been *designed* by the evolutionary process to learn, and how have they been designed to do it?

It is only by acknowledging that learning and adaptations are not separate things that we can begin to ask questions about how learning capacities are designed. Indeed, I will argue, our main questions about development should not be about innateness per se. Instead, they should be questions about developmental *design*. How are developmental systems designed by natural selection to enable them to produce adaptive phenotypic outcomes, and in what way are they designed to use environmental input to do so? In other words, in what ways do learning mechanisms exhibit mind-world fit?

Finally, there is the problem of "novelty." It's a fact about natural selection that it is a process that operates over evolutionary timescales. Some of these are short and some are long. Some adaptations have been crafted over billions of years, some millions, some thousands, or less; it depends on the case. But it's nevertheless a fact that all of an organism's adaptations were engineered in the past. This means that natural selection cannot have engineered adaptations to things that didn't exist until right now. It can't foresee the future. To some, this implies the following conclusion: Because our "specialized" adaptations are specialized to the *past*, they can't explain our capacity to deal adaptively with evolutionary novelty, things that didn't exist until the present. Instead, according to this view, we need to invoke adaptations that are not specialized: general-purpose adaptations.

Ultimately, I'll argue, this too is misleading, for the same reasons given above: If we handle novelty adaptively, then there must be evolutionary design that is allowing us to do so, and it has to come from somewhere. After all, rocks and sand don't handle novelty adaptively. It is only because we have brains with certain design

features—presumably rather elaborate ones that many other animals don't have—that we are able to respond in the way that we do to a changing world, a world changing largely because of us. The answer will involve learning and culture, without a doubt. But it will also involve adaptations with design features that *enable* those things: that enable us to have culture, that enable us to learn from others, and that enable culture itself to evolve. This implies specialized design. As for everything else about the mind, the answer to why we can do what we do in modern environments must involve mind-world fit, and we will need to explain exactly how that fit arises in evolutionary terms.

In this book I'd like to build a case for what I think a properly "holistic" evolutionary psychology should look like: an evolutionary psychology that brings all mental phenomena, from brain development to culture to consciousness, under the rubric of evolutionary explanation—at least potentially. But let me emphasize that my intention is not to give a complete account of how the mind works. Nor will I claim that all of the phenomena I'm talking about are completely, or even mostly, understood. Far from it. Although evolutionary psychology has made substantial contributions to understanding the mind over the past two decades, it is still in its infancy. There is no question that as we discover more about the brain, aided by accelerated technological advances in areas like genetics and brain mapping, the theories and methods of evolutionary psychology will have to evolve. However, I think we are already in a position to see what *kind* of framework we'll need for thinking about how minds evolve to fit the world. That is the kind of framework I aim to depict here.

The book is organized thematically into five parts. However, in keeping with one of the book's main themes—interaction—these parts build on each other and are thus not independent. In part I, "Evolution," I will begin with the theoretical foundations we'll need to understand the evolutionary forces that shape phenotypes—organismic design—and the special considerations we'll need to take into account when thinking about mental phenotypes, which evolve in a world of information. In part II, "Information," I turn to how the evolutionary process shapes mental mechanisms, beginning with mechanisms of perception and extending into the more complex mechanisms underlying human thought. Rather than attempting exhaustive coverage, my approach will be to use case studies from different areas of cognition, showing how they can be examined under the evolutionary lens. In part III, "Development," I turn to the crucial issue of development, and how evolutionary processes shape developmental systems to make them both flexible and capable of achieving adaptive outcomes. Part IV, "Culture," uses this view of development to understand how culture and the human capacity to deal with novelty can be conceptualized within the framework of specialized mental adaptations. Part V, "Architecture," addresses questions of how the mind is organized at the global level, how it can be both massively

specialized and yet massively interactive and interdependent, and how this complex whole evolves over time.

My goal is primarily a positive one: to depict, in broad strokes, what a whole-brain evolutionary psychology might look like. But because there has been so much confusion and debate in this area, I also hope to dispel some misunderstandings along the way. These include some of the misunderstandings I've mentioned above: for example, that it's possible to neatly sort adaptations into two types, special-purpose and general-purpose, or that if adaptations evolved in the past, they can't deal flexibly with the present. But I also hope to dispel one of the biggest misunderstandings of all: the idea that evolutionary psychology itself is some sort of "hypothesis" that can be falsified or confirmed. When done properly, evolutionary psychology is just evolutionary biology applied to the mind. Short of creationists, nobody would take evolutionary biology itself as a hypothesis. Like any field, evolutionary psychology contains theories and ideas that may or may not prove to be correct, just as does economics, or anthropology, or other branches of psychology. Unfortunately, the rise of evolutionary psychology has led to rampant generation of evolution-flavored hypotheses throughout the social sciences, not all of which pass basic tests of evolutionary plausibility. One of my goals is to show you that not all evolutionary hypotheses are created equal, and that careful thinking about the details of how evolution actually works—appropriately tailored to the circumstances in accordance with the first law—can get you a long way.

It's a sad fact that the evolutionary social sciences have gone through a series of fads, with various schools, like sociobiology, appearing and disappearing. Ultimately, I'd like to convince you that while any particular claim or theory in evolutionary psychology may prove to be false—that is what makes it science—the fact that human minds are products of evolution is something that we can be as sure of as any other fact in biology. This means that an evolutionary psychology of some kind is inevitable. But of all the evolutionary psychologies we might have, our shared goal, I hope, is to hit upon the right one—a description of the evolution of the mind that is, within the boundaries of what it's possible to know, the most likely to be correct. In order to do this, we must be relentless in remembering that what we are studying when we are studying the mind is not a computer, or a network graph, or a set of equations, or any other human artifact. It is a product of biology, and ultimately obeys no laws other than biological ones. What I hope to sketch for you, then, is an evolutionary psychology grounded as rigorously as possible in the logic of biology—the logic of evolution—and no other.

A final note about style and readership: This is intended mostly as a scholarly book, not a "trade" or "pop" book. I will, therefore, assume some familiarity with the basics of biology and social science. However (and perhaps in a break

with scholarly tradition), I have also tried to make it readable. To do this, I have adopted a somewhat informal tone, addressing the reader as "you" and presuming at times to teach you something. However, I recognize that you, reader, may know much of what I am saying already, so I ask your indulgence for this stylistic technique. I also hope that the book can be read by curious readers with little or no background in evolution or social science, if they are willing to go the extra mile to understand terms or concepts that I haven't explained. If there is a technical term I haven't defined, I'm intending whatever the standard reading of the term is in the literature, so an Internet search (with care taken to consult reliable sources) should suffice to figure it out. However, in those cases where I provide my own explanation or definition of a term or concept, please assume that I intend *that* meaning and no other. In my view, sloppy use of language and concepts is largely what's gotten us into the mess I described above, and I am hoping to remedy that, at least in part, by being careful about words.

PART I

# Evolution

# 1

## ADDITIVISM

What I am going to argue in this book is that evolution has built biological machines with a unique function: to create representations of the world, and to use these representations to navigate within it. The central premise of the book is that in order to understand what these machines do and the representations they create, we need to understand why they are here.

The reason that these biological machines—minds—are here is because they helped us to survive and reproduce over evolutionary time. We didn't necessarily need them—plants seem to do just fine without them (they do make decisions, but on much slower timescales). The reason minds appeared in animals seems to have something to do with the fact that animals move around. We don't merely absorb nutrients from the soil or air; we go get food and eat it. We don't sit around waiting for pollen to land on us in order to reproduce; we go in search of mates. And we don't fend off predators by tasting bad; we run away from them. We need *abilities* to do these things. What minds appear to do seems to be more or less as follows: they (1) allow us to identify things in the world (objects, structures, patterns, including other people and what is in their minds); (2) tell us what attitudes and goals to have with respect to them; and (3) move our bodies about in ways that are appropriate to those goals.

I'm sure you represent yourself and your brain as doing much more than that. You do things like go see movies, think thoughts that do not end up moving your body around in any way, and read books like this, none of which seems relevant to survival and reproduction. Nevertheless, I am going to claim that the reasons you can do all of those things can ultimately be tied to the three categories above. While it's certainly possible to think thoughts that have no consequences for survival and reproduction, it's pretty much an entailment of the way evolution works that the fancy machinery that allows you to have thoughts with no survival consequences only evolved because of the ones that did. It is in that sense that I will claim that every thought you are capable of having is the *kind* of thought that helped us to survive and reproduce in the past, and bears, at some level, the signature of that reason for existing.

Let's start from the beginning. Way back in our past, as animals started to be the kinds of organisms that moved around, things like finding food and mates and escaping predators were only achievable because of skills, albeit crude, of detection, analysis, and decision. This was why what we call "cognition" appeared. And the evolution of the brain and mind was, like everything else in evolution, a hill-climbing process, with later innovations building upon earlier ones. At first there was no cognition at all: Organisms could not detect anything, could not make any decisions, did not have any attitudes, preferences, goals, or knowledge, and could not act. Then early cognition appeared, based on very simple discriminations and decisions: light versus dark, cold versus warm, move away from, move toward. As time went on, things got more complicated. Dark things blocking out the light became discriminable as things that moved versus things that didn't. The ability to make more nuanced distinctions between things, and more nuanced decisions about them, appeared: telling one's own species from others, discriminating friend from foe, deciding when to eat, when to move, when to flee, when to hide. In each case, new innovations built upon older ones, modifying and adding onto what came before. Older, cruder abilities became the baseline onto which finer and finer grained capacities could be built—not necessarily throwing away the old skills, but improving on them, adding more.

What these examples and the many others we explore in this book point to is a general evolutionary principle: that the evolution of increasing cognitive complexity and flexibility—the ability to respond in a more and more fine-grained and context-sensitive way to the world—is generally (though not always) the result of natural selection *adding* new information-processing abilities to organisms, not taking them away.

This is not to say that all fitness-improving changes occur via addition of new things. The currency of natural selection is survival and reproduction, and when those are improved by removing something, natural selection does it. Humans, for example, have lost the ability to digest certain kinds of foods, smell certain kinds of compounds, and synthesize certain vitamins that our ancestors could digest, smell, and synthesize (Gilad et al., 2003; Milton, 1999). Each of these losses occurred because the fitness benefits of the ability in question were not sufficient to counteract the design-erasing effects of mutation—a principle that biologist Sean Carroll refers to as "use it or lose it" (Carroll, 2006). But loss of skills is not likely to be the primary explanation for the accumulation of the kind of cognitive complexity we would like to explain in humans: the kind that makes us more flexible, able to do a wider variety of things than any other species can. No matter how you slice it, making more of this has to be done by getting more information into the decision-making system. As we'll see, this can be done in myriad ways, exploiting many kinds of fit relationships between internal information and the external environment—but they all must be

accounted for in positive terms, in terms of what *enables* us to notice things, think things, do things, and decide things that other species cannot.

This is the view that I will call "additivism." You might be surprised that I feel the need to give it a name, or even to point it out, so obvious might it seem. However, the way that many people think about the role that natural selection plays in shaping cognition suggest that the skill-adding role of selection is not obvious at all. I will contrast it with another view, which is widespread in psychology and the behavioral sciences more generally, that I will call "subtractivism." This view is that the role of natural selection is fundamentally to *constrain* systems in useful ways: to narrow down the set of possibilities of what they could otherwise do.

If you do a literature search, you won't find any psychologists explicitly calling themselves subtractivists. I made the term up. What I am trying to describe is a point of view that, while acknowledging that evolution is important in explaining mind design, views its role as mainly one of adding constraints to otherwise general-purpose learning systems. For example, many neuroscientists and psychologists view brain development as resulting from the domain-general learning properties of neural networks filtered through biological constraints that point development in certain directions (Elman et al., 1996; Karmiloff-Smith, 1992, 1998; Quartz & Sejnowski, 1997). Somewhat differently, there is a school of thought in developmental psychology that holds that infants are endowed with a set of innate "theories" at birth, which are then revised by a set of general-purpose theory-revision mechanisms that alter the theories on the basis of experience (sometimes called "starting state nativism," Gopnik & Meltzoff, 1997; Carey, 2009). In general, it is common in developmental psychology to view cognitive development as proceeding via constraints on learning (Hatano & Inagaki, 2000; Karmiloff-Smith, 1992, 1998; Keil, 1981; Saffran, 2003).

Depending on how these proposals are operationalized, they could be viewed as painting a rather stark and minimalist picture of the role evolution plays in shaping developmental outcomes.[1] Pascal Boyer and I contrast what we call "ballistic" models of development with what one could call a "guided" view of development (Boyer & Barrett, 2005). In a ballistic model, developmental processes merely specify the starting state, aiming development in a particular direction and firing like a cannon. The projectile can be aimed at a target, but there is no systematic steering toward the target after launch. In contrast, one can imagine guided systems designed to hit a target through active steering—developmental outcomes like spoken language, social competence, or finding a mate. One can certainly think of both models in

[1] I'm not trying to argue that there isn't merit in aspects of these views or that constraints aren't important in development. What I'm arguing is that they aren't likely to be enough—more than just general-purpose mechanisms plus domain-specific constraints will be needed to explain the whole of cognition.

terms of constraints, and obviously, feedback from the world matters for both. But at least on the latter view, it makes just as much sense to say that development is targeted—designed to make some outcomes possible—as it does to say it is constrained, or designed to rule some outcomes out.

I would like to contrast these two views to show why it's important to think about evolved capacities as things that progressively add value—add information, add finer and finer grained distinctions, add abilities to react, and therefore add flexibility—to minds. This proposal is counterintuitive, and flies in the face of a more popular and common-sense view that sees flexibility and evolved mental structure as entailing a tradeoff, such that to the extent that a brain is flexible, it has fewer and fewer evolved constraints. By arguing against this, I won't be suggesting that the notion of "constraint" isn't important in evolution—it is. But in order to understand its proper meaning, we will have to see it as something that adds information to systems, rather than taking it away.

---

Let's start with an example. Imagine three different kinds of single-celled organisms: ones that can't move on their own, ones that can move but in random directions, and ones that can move in response to an environmental stimulus such as sunlight. All three such kinds of organisms exist. The world's earliest bacteria-like organisms were probably of the unmoving kind, bags of lipids containing innards of cellular machinery, floating around in ponds and oceans. Then, a series of evolutionary changes gave rise to structures called flagella: tiny whip-like biochemical structures that rotate like corkscrews. With these positioned on its body, a bacterium could move through the water. It could only move randomly, but this might be better than not moving at all. In ponds, for example, such a bacterium might have less of a chance of being stuck in an unfavorable spot or using up the resources in its local area. Reasons like these are why the changes in the bacteria's DNA that gave rise to flagella were retained. In this case, it's clear that natural selection created an ability where none existed before: the ability to move (Mitchell, 2007).

Once mobility is present, this sets the stage for further evolutionary changes that refine the mobility, making it responsive to particular environmental events. One such change, which seems to have occurred more than once in the history of life, is the transition to being sensitive to light, or what biologists call phototaxis (Jékely, 2009). Phototaxis can be thought of as a kind of cognition, or information-processing. It doesn't necessarily involve a brain, though of course it can. Like all cognition, phototaxis involves detecting a "cue," or information—in this case, sunlight—and then reacting in an appropriate or "adaptive" manner on the basis of it. Biologists

distinguish two kinds of phototaxis, positive and negative: approach and avoidance of light, respectively. Light can be good for things like photosynthesis and bad because of things like ultraviolet radiation. Single-celled organisms such as bacteria and photosynthetic algae exhibit, as you might imagine, diverse reactions to light depending on the usefulness or detriments of particular wavelengths and intensities of light for survival and reproduction, sometimes swimming toward light, sometimes swimming away, and sometimes exhibiting both reactions in a context-dependent fashion in a single organism (Braatsch & Klug, 2004; Foster & Smyth, 1980).

In order to get from the non-light-sensitive mobility afforded by flagella alone to the more complex or reactive mobility seen in phototaxis, an evolutionary change is required. In particular, what is required is the appearance of what are called sensory transducers: physical mechanisms, such as assemblages of molecules in the cell membrane, that are capable of detecting environmental events—in this case, light striking the surface of the bacterium—and causing a change in the behavior of the flagella, thereby altering the motion trajectory of the whole bacterium. In modern phototactic bacteria, this transduction stage involves several steps, each of which had to have arisen for a reason, offering some fitness benefit to the organism (Bardy et al., 2003; Ng et al., 2006). For our purposes I'll consider them as a unit, because what matters for us are the functional consequences of the appearance of the sensory transduction mechanism: what it means for the behavior of the organism and *why* this change would be selected for.

In a nutshell, what gets "added on" to bacteria in order to make them respond to light is a biochemical signaling pathway that links a chemical light detector called an opsin, via a series of chemical steps known as a signaling cascade, to the behavior of flagella. An opsin is a kind of protein (a complex molecule made from many amino acids, the sequence of which is encoded in the bacterium's DNA) that changes its conformation, its physical shape, in response to certain wavelengths of light (Goldsmith, 1990; Terakita, 2005). In the phototaxis of the bacterium *Natronobacterium pharaonis*, for example, light hitting opsins in the cell membrane causes a kind of lever, a coil made of protein, to extend into the cell and flip a switch via contact with another kind of molecule called a kinase. The kinase phosphorylates (adds a compound known as a phosphate group to) yet another kinase that acts as a switch for the flagellar motor. There are thus two switches, or transduction stages: one where light is tranduced via the opsin into a chemical signal, and another in which that signal turns on the flagellar motor (Gordeliy et al., 2002).

This is all very detailed, and we needn't worry too much about all the details here. In fact, adding to the complexity, bacteria such as *N. pharaonis* are able to regulate or adjust their reaction to light via other chemical pathways that adjust the signaling cascade's sensitivity. And of course, the machinery that we see now in modern-day

*N. pharaonis* is the result of literally billions of years of evolution, and therefore likely to be substantially improved over whatever the first step was that made its ancient ancestors responsive to light. But let's ignore all that for now and consider the consequences of the most basic, initial, evolutionary addition: the insertion into the membrane of a chemical sensitive to light, which, in whatever way, stimulated flagellar activity.

Note that there are various ways in which positive or negative phototaxis could be produced. For example, putting an opsin on the opposite side of the cell from where the flagellum was located—as if on the bow of a boat—would cause positive phototaxis, because it would cause the flagellum to push the bacterium toward the front when light was detected there. Putting an opsin right next to the propeller, on the other hand, would cause the propeller to push away from light whenever it was detected at the "stern," resulting in negative phototaxis, or light-fleeing (Braitenberg, 1984; Foster & Smyth, 1980).[2] Let's imagine, for the sake of example, the latter case: a mutation causes light-sensitive opsins to appear adjacent to flagella, resulting in negative phototaxis.

Of course, such a mutation will only be favored in environments in which moving away from light is a good thing for the organism's survival and reproduction, such as an environment where ultraviolet light is a significant source of mortality. Imagine bacteria getting very near the edge of a shallow pond, where sunlight will be extremely intense and there is even a possibility of drying up. In such an environment, moving into deeper waters—anywhere away from the edge—is a good thing. If this change in organismic design results in a net benefit for the bacteria, then it will spread in the population, and natural selection will have added an ability where previously there was none: the ability to respond adaptively to light.

What "adaptively" means, here, is that the response is of a particular kind. It is systematically the response that tends to increase fitness—survival and reproduction—relative to other possible designs, or ways of reacting or not reacting to light. In an environment where UV light poses significant mortality, bacteria with light-sensitive opsins next to their flagella will run away from bright sunlight, thereby surviving to leave more copies of themselves than bacteria without them. This is because the phototactic bacteria are more reactively flexible than the non-phototactic bacteria,

[2] In some species, such as the phototactic algae *Chlamydomonas*, the body of the organism rotates so that the moving eyespot creates a kind of directional antenna—like a satellite dish that samples signals from multiple angles to triangulate the direction of light (Foster & Smyth, 1980). Note too that sometimes flagella can be mounted on the "back" of the organism, like a tail, and sometimes, as in *Chlamydomonas*, they are mounted on the front and sweep backward, a bit like oars.

adjusting their behavior in response to circumstances in a way that the light-blind bacteria cannot.

Has natural selection "constrained" the bacterium's behavior in this case? In a sense you might say yes, because previously the bacteria could go in any direction, and now their movement is channeled away from the light. Indeed, the behavior exhibited by phototactic bacteria is sometimes called a "biased random walk," because it still exhibits lots of randomness, but randomness that is constrained to tend away from the light.

This is a perfectly legitimate use of the word "constraint," and we will discuss the various meanings of the word in more detail below. But it should not obscure the larger point that in this example evolution has added a new capacity, not made some previous capacity weaker or less flexible. This does indeed entail constraint on at least two levels: The bacterium's design has been constrained by evolution to a particular part of possibility space, and the phototactic mechanism itself operates by constraining the bacterium, or biasing it, toward one direction with respect to light rather than the other. But, importantly, information has been added to the system: information about light and dark, and how to respond to them. This information, of course, is not stored as any kind of explicit representation, but is rather implicit in the structure, or the design, of the organism's adaptations (the opsins, flagella, and the signaling pathways between them).

In fact, this is a common feature of the way form-function fit is engineered into behavior-guidance systems. What in some ways looks like "knowledge" of the environment—something akin to a knowledge that "UV light is bad," or an if-then rule that says "when you detect UV light, move away from it"—is engineered implicitly into the structure of the system. When I say "implicit," I mean that the information is not present in the form of a representation or message that can be read, like a sign on the fridge that says "Remember to put on sunblock." Instead, the system is engineered such that the bacterium simply moves the right way under the right circumstances. By "right," I mean "fitness-right"—a terminology that I will use throughout the book to replace a more cumbersome formulation that would be something like "in ways that led to higher survival and reproduction than other ways did, aggregated over past evolutionary time and space."

The example of phototaxis illustrates a concept that will play a central role in this book: the concept of an *inductive bet*. There is a sense in which negatively phototactic bacteria appear to be betting that sunlight is bad, by systematically moving away from it. But there is no conscious thought behind the bet, or indeed any mind at all in the sense that we typically use the term. Instead, the bet is embodied in the design of the organism's adaptations, in the form of physical features that cause it to move away from light. In fact, this is a general feature—one of the few, as the first law

of adaptationism suggests—of adaptations: They embody statistical "guesses" about what their environment is like, and by extension, what it will be like ten minutes from now, and tomorrow, and the next day. The design features of adaptations are the way that the statistical number-crunching process of natural selection, acting over evolutionary time, transmits the results of its number-crunching to the present: in the form of actual stuff, causal stuff like opsins and flagella that act in certain ways when placed in the world.

In this sense, we can think of the products of natural selection, such as opsins and flagella and signaling pathways, as embodying strategies for dealing with the world. What I am calling "mind-world fit" is the fit between those strategies, those inductive bets, and the world. A theory of inductive bets allows us to generate hypotheses about the circumstances in which those bets will pay off: namely, circumstances in which the causal properties of the world on which the bets rely, those properties that played a role in their evolution, continue to hold true. Adapting a term from the philosopher J. L. Austin, I will call these the *felicity conditions* of the bet: conditions that, if they hold true, lead to an inductive bet paying off (Austin, 1962).[3]

Technically, induction is a form of inference that involves generalizing, under uncertainty, from specific cases (Goodman, 1954; Holland et al., 1986; Popper, 1959; Quine, 1969; Tenenbaum et al., 2006). When you guess that a dog that you've never seen before can bark based on having heard other dogs bark, you're making an inductive inference, because you don't actually know for sure that this new dog can bark. The same goes for your expectation, or hope, that the sun will come up tomorrow: It's not a *deductive* certainty, because it lies in the future and depends on many physical states of affairs continuing to hold true. Nothing guarantees it as inevitable, though it is a pretty good bet. Such bets, of course, needn't be all-or-none, but can be probabilistic, with some probability of paying off. In this case, we can think of the design of the opsin-flagellum signaling system I've described as betting that UV light is bad, and that moving away from it is a good idea. Whether or not the bet ends up paying

[3] Austin proposed the concept of felicity conditions as part of his theory of speech acts. Speech acts include things like requests ("Can you please pass the salt?") and declarations ("I promise to pay you back next week"). Austin noted that it doesn't really make sense to call these kinds of utterances "true" or "false." Whereas a statement like "It's raining outside" has truth conditions—it needs to be raining in order for the statement to be true—things like requests and promises don't have truth conditions. Instead they have felicity conditions, things that make them, in essence, work. For example, a felicity condition for "Please pass the salt" is that there must be someone within hearing range who can pass the salt. Otherwise, the request makes no sense and will fail. Here, of course, I am using the term slightly differently, and it is therefore not entirely analogous to felicity conditions in speech acts. The similarity is that felicity conditions for inductive bets are things that need to hold true in order for the bet to pay off.

off depends on it being the case that UV light actually *is* fitness-bad in the present circumstances, and that moving away from it is a fitness-good idea. In that case, the bet turns out to be felicitous, or adaptive: It increases fitness, relative to other options.

The probabilistic goodness or badness of inductive bets corresponds to the principle sometimes known as *ecological rationality*: What makes any strategy rational is the fit between that strategy and the structure of the world (Brunswik, 1955; Gibson, 1979; Gigerenzer et al., 1999; Simon, 1990). Importantly, in most cases, including almost all cases of biological adaptations, the nature of the fit between the strategy and the properties of the world is a statistical one. Evolved inductive bets are not guaranteed to work out, but rather, are bets that have survived the process of natural selection. In fact, the stochastic nature of the world means that even in the environments where the mechanism evolved—what evolutionary psychologists call its "environment of evolutionary adaptedness," or EEA (Bowlby, 1969; Tooby & Cosmides, 1992), also sometimes called its "species-typical environment" or "normal environment" (Bjorklund et al., 2007; Morton & Johnson, 1991; Samuels, 2004)—it will be wrong, is *guaranteed* to be wrong some of the time. For example, some bacteria that are doing just what their opsins and flagella evolved to do will nevertheless die because of it. How often, when, and why are matters of the first law.

Of course, in the case of opsins and flagella, no actual inference is occurring. I'm using a metaphor. In the metaphor, the experience of the organism, akin to your experience of dogs barking and suns rising, is the summed evolutionary experience of the organism's ancestors interacting with their environments plus what the organism has learned in its lifetime. The first kind of experience is what we will call the *history of selection* that has shaped the organism. The second kind of experience, as we will see, shapes organisms via the developmental mechanisms that the first kind of experience, the history of selection, has built. What organisms are doing when they make inductive bets is to leverage or exploit those properties of environments that have proven to be reliable, or at least statistically expectable, in the environments where their design has been shaped. This, ultimately, is what we will mean by mind-world fit.

As we go along, I will be developing this metaphor of inductive bets as a way of resolving what some see as a kind of evolutionary paradox, which we might call the "paradox of novelty": If the design features of organisms have only been shaped by past events—which seems a causal certainty—then how can organisms respond adaptively to *current* events, and by extension, how (and when) can we expect them to react adaptively to events that haven't happened yet, like the sun rising tomorrow? The answer, I will suggest, lies in mind-world fit. When organisms are able to respond adaptively to novelty, it is because at least some of their inductive bets are warranted by the causal consistency of the world. In other words, organisms respond adaptively

to novel circumstances when they are felicitous, either by chance or by design. This is the key to understanding how, when, and why adaptations work.[4]

---

This way of thinking about adaptations as involving bets, or strategies, might seem quite intuitive, but I'll argue that it requires altering our intuitions about several other concepts that are intimately related to the concept of adaptation: in particular, the concepts of specialization and innateness.

Every adaptation is specialized. It's just a question of *how*, or in what way. When stated in this way, this idea seems uncontroversial. However, it does fly in the face of standard use of terms like "specialized" and "domain-specific" as delineators of *categories* of mechanisms or processes in the mind. Imagine for a moment that this categorization scheme is legitimate, and that we really can divide mechanisms into the domain-specific ones and the domain-general ones. Which kind are opsins and flagella?

Clearly, they are specialized. They carry out specific functions, and they use specific kinds of information to do so. We'll return to the concept of "information" in more detail below, but what I mean in brief is that they interact with specific properties of internal and external environments. Opsins react to information in the form of light waves—and not just any light waves, but particular regions of the frequency spectrum, such as UV light. This domain specificity is embodied in their design in that the changes in conformation caused by exposure to UV light result from specific properties of the protein molecule that have been tuned by evolution to change shape in particular ways as a result of the energy contained in those frequencies (Goldsmith, 1990; Shi & Yokoyama, 2003). The same goes for the components of the signaling cascade mediated by the kinases that phosphorylate the flagellar motor. Each responds to very specific "inputs" or events that, as we'll see when we discuss the technical definition of information, carry information of a very specific kind: Namely, that UV light has been detected in a region of the bacterium's membrane. In each case, the mechanism is domain-specific in that it doesn't respond to just *any* information, but information of a specific kind.

[4] How could novel circumstances be felicitous "by design"? In some cases, circumstances that appear novel may in fact be non-novel along some dimension or dimensions that satisfy an adaptation's inductive bets; thus, in a sense, the adaptation has in fact been designed for those aspects of the circumstance. In other cases, aspects of the world can themselves evolve to satisfy the inductive bets of evolved mechanisms. As we will see later, the products of culture, such as human-made artifacts, often evolve to be felicitous in this way.

You might argue that domain specificity shouldn't really be regarded as a category, but as a dimension. On this view, domain specificity is a matter of degree: That is, while there is no such thing as a truly "domain-general" mechanism, it's nevertheless the case that mechanisms vary in how domain-specific they are. This would mean that mechanisms can be lined up on a kind of continuum of specificity, with some lying on one end of it, others at the other end, and still others arrayed in between.

Perhaps that's a little better than thinking about specialization in a binary, either/or way, but I think it's still not quite right. It *is* true that domain specificity is not just a single thing, and that the nature of domain specificity, the nature of fit between mechanisms and information, varies widely across all the different information-processing mechanisms that the evolutionary process has created. Indeed, this is an entailment of the first law. But the problem with thinking in terms of a specific-to-general continuum is the implication that there is just a single dimension of specificity. Instead, the space of possible matches of mechanisms to information is as massively multidimensional as the kinds of information there are in the world, and in minds.

What this means is that we can't, for example, order opsins with respect to kinases in terms of how "specific" or "general" they are. They just respond to different kinds of information: stimulation by light and stimulation by molecular contact, respectively. It might be possible to salvage a continuum view of specificity-to-generality *if* we restrict its use to cases where two mechanisms use exactly the same kind of information but one mechanism uses a narrower portion of it—but as we'll see, that really doesn't apply to most of the differences in informational specialization that we'll be talking about in brain mechanisms.

On the view of domain specificity I'm advocating, then, *all* adaptations that have evolved to process information, from perceptual mechanisms to reasoning mechanisms to learning mechanisms, have a domain (Barrett, 2009). This is true in at least two senses. Anthropologist Dan Sperber introduced the important distinction between an adaptation's *proper* domain—the set of conditions in which it was designed to operate, or inputs it was designed to process—and its *actual* domain, the set of conditions in which it can, in fact, operate, whether designed to do so or not (Sperber, 1994). Proper domains are defined in terms of evolutionary history, and actual domains are defined in terms of an adaptation's design features and the scope of things they enable the adaptation to do. By analogy, the proper domain of a hammer is nails, but it can be used to hammer things other than nails, and even to do things other than hammering (e.g., serving as a paperweight is within a hammer's actual domain). The point is that all adaptations are domain-specific in both senses: They have a set of things they evolved to do and a set of things they *can* do by virtue of their design. In order to understand information-processing mechanisms and how they work, then, it isn't particularly helpful to label them as "specific" or

"general." What is helpful is to provide a positive description of the domain of information a given mechanism evolved to operate on, and how its design features fit that domain. This leads to hypotheses about mind-world fit that we can test.

---

What about innateness? You might be completely on board with the example of phototactic bacteria and what it shows about adaptation, specialization, and domain specificity. But, you might point out, it's a far cry between bacteria, which don't really have "minds" in any sense that we'd typically use that term, and humans, who do. In the bacterial case, you might argue, the behavior-regulation mechanisms are about as close to being "hardwired" into the DNA as you can get. Humans, on the other hand, have things like learning. In the human case, how will we ever know which aspects of mind-world fit come from evolution, and which come from learning, experience, culture, and the like?

Development is perhaps the single biggest sticking point in contemporary debates about the mind. Often the disagreement is not about what adult minds are like (though there is still plenty of debate there), but rather, how and why they get that way. Traditionally, you've got the empiricists, who tend to focus on the role of learning and experience in shaping the mind, and nativists, who tend to focus on genes and innateness. There is no question that both are part of the answer, but one of the things I'll be arguing in this book is that we're probably going to need a toolkit that includes more than just genes and learning. I'll also suggest—in line with many critics of evolutionary psychology—that the intuitive concept of "innateness" is often not up to the job of explaining what we want to explain (Elman et al., 1996; Griffiths, 2002; Griffiths & Gray, 1994; Griffiths et al., 2009; Lickliter & Honeycutt, 2003). If innateness means presence at birth, no role of experience in constructing the trait, and/or strict uniformity across individuals, probably few traits fit the bill. On the other hand, this doesn't mean that natural selection can't "specify" traits, as some would suggest. While it might rarely specify point outcomes, natural selection must have a systematic effect on phenotypes if natural selection occurs. This means that we're going to have to develop ways of thinking about how natural selection shapes design that are consistent with what contemporary biology tells us about how development works, and that are not reliant on a notion of innateness that holds that the effects of natural selection are limited to what's present at birth and that remains fixed thereafter. This idea of innateness does not come close to capturing the full scope of resources that natural selection has at its disposal for engineering adaptive design into organisms. In particular, it ignores the fact that organisms can alter their own properties, and that these alterations *themselves* involve inductive bets: They occur

because of design features, shaped by natural selection. This means that even features of organisms that are not "built in" show the hand of natural selection at work.

Many developmental psychologists would agree with this. Indeed, there is widespread agreement that natural selection operates by shaping the developmental systems that build organisms (Barrett, 2007; Griffiths & Gray, 1994; Lickliter & Honeycutt, 2003; Oyama, 2000; Tooby & Cosmides, 1992). However, many contemporary views of development in psychology continue to sneak in a dichotomy from the old nature/nurture debate: innate versus learned. They do this by conceptualizing development as being primarily caused by learning processes in which the information structures in the brain get built by two kinds of ingredients: learning mechanisms, which are relatively small in number and general in purpose, and constraints, provided by evolution—or sometimes by the structure of the inputs or the learning rules alone—which guide the learning process (Elman et al., 1996; Gopnik & Meltzoff, 1997).

There is certainly something to this idea. The brain is made of neurons, which have properties of self-adjustment. Brain development is a dynamic process in which neural structure self-assembles in interaction with internal and external environments, and natural selection must certainly provide the materials that guide this process. But there is also a sense in which this "two ingredient" view—learning plus constraints—seriously underplays the tools natural selection has at its disposal to build organisms, because it essentially limits all developmental processes to guided learning, or constrained searches through possibility space.

While learning plus constraints undoubtedly describes *some* of development, there is no reason to think it describes all or even most of it. For example, what we know about the development of body morphology suggests that the idea of a guided search through possibility space is a poor description of how most of the body develops (Carroll, 2005; Carroll et al., 2005; Raff, 1996, 2000). Instead, body development occurs through a progression of events in which contingent gene regulation—genes being turned on and off as cells divide and the body grows—create a series of spatial coordinates within what are known as "morphogenetic fields." These fields are themselves an internal environment that shapes subsequent bodily development, because the developmental system of genes and regulatory elements uses their position within this changing environment of spatial coordinates to decide what to do next. For example, in the fruit fly *Drosophila*, two coordinate systems, anteroposterior and dorsoventral, are established in the growing embryo via patterns of contingent gene expression. These are then further subdivided into a series of stripe-like domains that give rise to different tissues, which develop as a function of where they are in this biochemically structured environment, through differential gene regulation. Response of genes in the dorsoventral axis depends

on a gradient of expression of the so-called "Dorsal" (Dl) protein, regulated in part by the number of binding sites for the Dorsal protein in the regulatory elements controlling those genes. Many genes respond combinatorially to multiple activators or repressors in the gradient, leading to both greater specificity and diversity of developmental patterns in different parts of the body than if those genes did just one thing all the time. These events in turn drive further changes in the spatial coordinate system, which subdivides itself into progressively finer elements as the body grows (Carroll et al., 2005).

What all this means is that body development is highly dynamic, contingent, and context-sensitive. Genes don't merely "specify" outcomes in a one-to-one fashion; the outcomes they produce depend on where and when they are expressed and on prior developmental events. Nevertheless, body development is best described as a series of construction events that are *designed* to hit particular targets. Not point outcomes, of course, but regions of possibility space where the design features emerge through interaction—as they do everywhere in development—but at the same time emerge this way because of a history of selection to do so. What I will suggest is that this is likely to be a good description of brain development as well. However, it will require us to take into account that the kinds of targets brain development is designed to hit are not limbs and digits, but information structures. Whatever you might think of the "initial" state of the cortex, it's clear that the *developed* cortex of the brain is highly domain-specific, rather than uniform. This is not to say that cortical regions are uniquely responsible for doing different things, but there is no question that parts of the cortex vary in their involvement in different specialized processes: speech, memory, recognizing objects, navigating space, and many more, at many scales of specificity (Bilder et al., 2009; Poldrack et al., 2009). There is no reason why these patterns can't emerge developmentally as both a result of experience *and* contingent patterns of development—including, perhaps, contingent gene expression—rather than a single all-purpose "plasticity" mechanism. We will return to this idea of brain development as analogous to contingent processes of body development later on.

What I am going to argue, then, is that much of brain development is best seen as an active process that is designed to produce certain kinds of outcomes. Learning is crucial to this process, but constraints on learning are only one way, and perhaps not even the most important way, that natural selection instantiates "design." This argument will take me most of the book to develop. But let's begin by thinking about what natural selection does when it alters organismic design, and what this tells us about the additive and subtractive aspects of development.

---

Constraints can be seen in a glass-half-full and a glass-half-empty way. Most literally, of course, constraint implies prevention. If I lock you in a room, I prevent you from leaving and thereby constrain your movement to the inside of the room. Constraints are causal: They entail not merely the removal of something but the addition of something that can have causal effects on the future states of a system, such as the effects of a locked door on your possibilities of movement. Typically, when the word "constraint" is used, it refers to a narrowing down of a set of possible outcomes rather than exact determination of a point outcome—though in extreme cases it can. For example, if I duct-tape you to a chair bolted to the floor, I constrain the spatial system of you and the room to a point outcome, entirely determining your location. We could similarly speak of just about any active process of causal guidance in terms of constraint if we wanted to; for example, we could say that a pilot flying from Los Angeles to New York "constrains" his plane to land at JFK airport via his actions.

It would be a bit odd to use the word "constrain" in this way, of course (for one, "constraints" are usually seen as unalterable and exogenous to the system in question, e.g., the decision-maker). But these examples illustrate the glass-half-full/glass-half- empty nature of the language of constraint and causation. Causation involves both possibilities foreclosed, or what doesn't happen, and possibilities generated, or what does. In this sense, the causal properties of all material stuff in the world are therefore both constrainers and enablers, and in any given case, we can think and talk about both aspects of causation. In other words, we can talk about what the causal properties of stuff enable, and what they rule out. Developmental psychologists Rochel Gelman and Earl Williams have introduced the concept of an "enabling constraint" to capture just this dual nature of constraints in causation (Gelman & Williams, 1998).

It is this distinction that I have in mind when I talk about "additivism" and "subtractivism." Biological systems are made of material stuff and are therefore causal. Natural selection shapes the properties of that stuff because of the causal effects that it has in the feedback loop of reproduction: Stuff that is causally fitness-better tends to win out, in the long run, over stuff that is causally fitness-worse. Like any causal stuff, we can think about the properties of biological systems in terms of what they enable and what they constrain. Both aspects of causation are important, as we'll see. But what I will argue is that thinking about the enabling functions of adaptations, what kinds of outcomes they make possible by virtue of their properties, is going to be critical for understanding the properties of minds that we want to explain, such as their capacities to form representations of the world, to make inferences, to learn, and to be flexible. And to understand what kinds of outcomes adaptations enable, we're going to have to think about why they were favored by selection by being explicit about how their causal properties, in interaction with the world, improved fitness.

This is true of development as well, and is the key step toward understanding developmental design without getting caught in the trap of the folk notion of innateness.

To do this, however, we will have to think carefully about the evolutionary process, the developmental process, and the relationship between them. Both of these have been likened to "hill-climbing" processes because both are anti-entropic, creating states of increasing order out of disorder. However, the mechanisms whereby this hill-climbing occurs are not the same in the evolutionary and developmental processes. In particular, natural selection acts as a kind of passive filter that retains fitness-better variants, but by itself—in the absence of mechanisms that cause organisms to grow and reproduce—it does not do anything, does not generate anything. The mechanisms of development, on the other hand, have been selected precisely because of their active, outcome-generating properties. There are important senses in which we can talk about "constraint" in both cases, but the mechanistic details of what constraint means are not at all the same in evolution and development.

There is no question that the processes of evolution and development are intimately linked, so much so that at some level it is impossible to separate them. Evolution is just the process of development looped over and over, with the products of development—organisms—competing with each other. We can therefore say that natural selection shapes development, but that is not the end of it. As the field of evolutionary developmental biology has increasingly made us aware, the reverse is true as well. This is because, at heart, it is the products of development that natural selection acts on. Although many evolutionary changes originate in mutations in DNA and epigenetic material, it is developmental processes that generate phenotypes, which are the vehicles of survival and reproduction. Variation in phenotypes is what causes variation in fitness—differential survival and reproduction—which is natural selection. Because past evolutionary events are responsible for the developmental systems that exist and the variations that they can generate, it is past evolutionary events *and* the current environment—not just the environment alone—that shape the course of evolutionary history.

Evolutionary developmental biology, or "evo-devo," explores how the regulatory systems that control gene expression generate the development of phenotypes, and how these developmental systems are shaped by processes of descent with modification over time (Breuker et al., 2006; Hall, 2003; Carroll, 2008; Raff, 1996, 2000). As we will see later, an understanding of these processes will be crucial for a complete understanding of mental mechanisms and how they evolve. Additivistically, evo-devo tells us what tweaks in development produce evolutionary innovations; understanding such tweaks is necessary for understanding how new functional stuff arises from old. It also leads to the important idea of an *evolutionary constraint* (Alberch, 1982; Brakefield, 2006; Gould, 1977; Maynard Smith et al., 1985). Evolutionary constraints

are, liked the locked door to the room that I mentioned above, constraints that limit the directions that evolution can go, defining the possibility space of options that can be generated by developmental systems, and therefore the designs that compete with each other at any given time.

Evolutionary biologists have developed a variety of taxonomies for categorizing evolutionary constraints, based on how they limit the variation on which natural selection can operate. Raff (1996), for example, distinguishes three kinds: physical constraints, genomic constraints, and constraints of organization, or complexity. The first kind surfaces in things like the physical properties of biochemicals of which organisms are made, like DNA, proteins, and lipids, and physical constraints in how these can be assembled in ways that make functioning organisms; limits on size and scaling of animal body parts are sometimes cited as examples. Genomic constraints, in Raff's taxonomy, include things like boundaries of genome size and the ability of new mutations to be generated by changes to existing genes. Finally, there are organizational constraints, or what might be called "you can't get there from here" constraints: cases in which organisms have been committed to certain forms of organization by the evolutionary process in ways that cannot be undone, in the sense that variants on those forms either don't appear or are not viable.

These constraints arise from what is sometimes known as *path dependence*: the dependence of current states of a system on events that occurred in the past, such that if those events had occurred differently, different things might be possible now. One consequence of evolutionary path dependence is what Maynard Smith and Szathmáry call *contingent irreversibility* (1995). Organisms and systems of organisms are complexly arranged such that the adaptive nature of many properties depends critically on *other* properties. This means that properties can become evolutionarily locked in because "undoing" one property (returning to an earlier state of the system, which might have been adaptive at the time) would depend on simultaneously undoing many others. This irreversibility can occur at many levels. Consider, for example, a mutation that suddenly changed the mapping properties of the genetic code in a human, or a mutation that caused a worker ant to strike out on its own. Although these mutations might have been advantageous had they happened earlier in evolution, they aren't now, because of the co-adapted nature of organisms' different design features. And of course, many of the contingently irreversible features of organisms may be "frozen accidents," arbitrary features that might have been otherwise if evolution had proceeded differently, but cannot now be reversed. Crick (1968) originally proposed this term to describe arbitrary but unchangeable features of the genetic code.

While we can distinguish between different causal mechanisms of evolutionary constraint, they are all in a sense forms of path dependence. Even physical constraints

arise because at some point in the past, organisms became committed to certain kinds of chemical building blocks. The same goes for constraints arising from the causal properties of developmental systems and what variants they can generate. And while evolutionary biologists sometimes categorize developmental constraints as a sub-category of evolutionary constraint (Alberch, 1982; Maynard Smith et al., 1985), all evolutionary constraints are in a sense developmental, since all arise from limitations on either what the developmental process can produce or the fitness-goodness of those products.

This path-dependence view of evolution—which Darwin himself recognized, calling it "descent with modification"—clearly fits with the picture of the stepwise evolution of behavioral complexity in organisms such as phototactic bacteria that I presented above. In such a system, each evolutionary change involves an alteration in design, and subsequent evolutionary changes occur because of variation in design that is already present. This is sometimes likened to a ratcheting process because at each step in the process, additional design is locked in, and the system is constrained to go forward, not backward, with respect to steps in the ratchet that have already occurred—at least usually. Why forward? If Maynard Smith and Szathmáry's idea of contingent irreversibility is right, the more steps one moves forward from a particular point in the ratchet, the harder it will be to return, evolutionarily, to that point (again, this is not a hard-and-fast rule but more of a general principle; exceptions can occur, such as horizontal transfer of blocks of genes in bacteria, jumping what would otherwise be multiple "steps" at once; Boto, 2010).

As Maynard Smith and Szathmáry point out, this doesn't necessarily mean that the evolutionary process is guaranteed to produce ever-increasing complexity.[5] But it does mean that newer adaptations take, as their starting point, the adaptations that already exist in the organism. Newer design variants exploit older ones. For example, in the evolution of phototaxis, either flagella existed, setting the stage for the evolution of light-sensing opsins that could make use of flagella, or opsins existed for some other reason, and flagella appeared as new design features that could exploit the light-sensing capacities of the opsin system. In the case of humans we will discuss many examples of such evolutionary "leveraging": cases in which newer adaptations only evolve because of adaptations already in place and whose design features therefore only make sense in light of the earlier adaptations. This is a kind of "standing on the shoulders of giants" in which no adaptation alone could do the work it does—or evolve—in the absence of other adaptations on whose work it relies. Opsins and

[5] See McShea & Brandon (2010) for an argument that increasing complexity should be the null model for evolutionary systems.

flagella show this in that the inductive bets of each rely on properties of the other. This could be seen as a kind of constraint if you like, but it is also a form of synergy, or distributed design. I will argue that thinking about such leveraging relationships between older and newer mechanisms will be critical to understanding human flexibility, and quite different from looking for the source of flexibility in single mechanisms alone.

---

This view of evolutionary hill-climbing turns the standard view of evolutionary constraints on its head. While hill-climbing can be seen subtractively as a source of constraints, it is also, additively, the source of complex design. The fact that new innovations evolve in organisms that already possess many other design features is how complex design accumulates. On the one hand, then, thinking about constraints is useful for thinking about designs that don't appear. But on the other hand, in order to understand the designs that do appear and why, it's useful to think about the *enabling* side of the accumulation of design.

This matters in development too. Some have pointed to the inherently subtractive nature of natural selection—the fact that it acts only as a sieve, pruning the variation that is there—as evidence that it does not "specify" phenotypic outcomes and is not itself a generator of design. It's true that natural selection at any given time is only a pruner and not generative. But for the reasons outlined above, natural selection, acting again and again on variation produced over time, *is* generative. And the developmental systems that it produces are generative as well: Because of the ratcheting process of cumulative selection, developmental systems become targeted to hit more and more specific developmental outcomes. Adult organisms are not merely "constrained" zygotes. Instead, the processes that build adults from zygotes are active, targeting certain areas of design space that would be hard to reach by a guided random walk. Guided random walks—trial-and-error plus correction—are good strategies when you can't foresee the region of possibility space where it is fitness-good to land, but not so good when you know where you're going. They are slow and inefficient ways of finding a target in possibility space, and in many cases, cannot be guaranteed to find the target in any finite amount of time. Unlike the evolutionary process, which has no foresight about future states of the system, the developmental process *can* have foresight. Natural selection acting over time can figure out what regions of possibility space—the space of all possible developed phenotypes—that it's best to land in, and engineer developmental mechanisms designed to hit those regions of the space, which evolutionary psychologists John Tooby and Leda Cosmides call *adaptive targets* (1992).

In psychology, the idea that organisms don't get to hard-to-reach areas of possibility space without help was perhaps most effectively introduced by the linguist Noam Chomsky (1965). There are many ways to arrive at the idea of "constraint" in this sense, but Chomsky got there by thinking about language learning. He thought about the problem of learning faced by a young child as a hill-climbing problem. Metaphorically, the child starts at the bottom of the "hill," having no knowledge of its language (I mean, here, the language being spoken around him/her, e.g., English in the case of children born into English-speaking homes). At the top of the hill, which he/she hasn't yet reached, is full adult competence in speaking English. The question that interested Chomsky was: What needs to be "built in" to get the child from the bottom of the hill to the top?

Note that in this metaphor, being at the bottom of the hill is easy. There are many ways to be at the bottom of the hill: babies, pigeons, and rocks are all there. As one goes up the hill, one proceeds toward states that are less and less random with respect to the target—knowledge of English—and therefore more and more difficult to achieve without help. Using a concept from physics, we could say that proceeding toward the top of the hill is proceeding toward higher and higher states of order.

Crucially, however, as Chomsky noticed, there are many possible hills of similar complexity, and only one of them would correspond to correct knowledge of *English*. The question he pondered was, starting at the bottom of the hill, and hearing some sample of English sentences being spoken around him (what he called the "stimulus" or the "input" to the learning system), what hill-climbing algorithm—what set of learning rules—would allow the child to get to the top of the correct English hill?

Chomsky proposed, controversially, that no general-purpose learning (hill-climbing) algorithm would be guaranteed to derive the correct grammar of English, because any finite stimulus—any finite set of English sentences the child might hear—is in fact consistent with an indefinitely large number of possible grammars. This problem is sometimes called the "poverty of the stimulus" (Chomsky, 1965; Pinker, 1979, 1984; see also Gold, 1967; Valiant, 1984, for mathematical arguments, and Laurence & Margolis, 2001, for a recent review). In fact, it is an example of a deep problem faced by learning systems, and one that goes under many names. The problem is sometimes called "opacity," meaning that what is to be learned is not transparent in the stimulus but opaque, or hidden (Csibra & Gergely, 2009). Others call it "underdetermination," i.e., the evidence is not sufficient to determine which of several hypotheses (e.g., grammars) is correct (Tesar & Smolensky, 2000). And in cognitive science, the "frame problem" refers to the fact that different ways of framing the same learning situation can lead to different answers or, sometimes, no answer at all (Dennett, 1984; McCarthy & Hayes, 1969). In all these cases, the problem is that

the necessary learning can't be achieved *just* by looking at the stimulus; something else needs to be added in order to narrow down the possibilities to the correct one.

To solve this problem, Chomsky postulated that the mind must constrain the possible grammars that it considers when trying to learn a language based on an input set of overheard utterances. Because children do, in fact, tend to deduce something like the correct English grammar when exposed to it in childhood, he suggested that we must infer the existence of some set of constraining mechanisms, which he called the "Language Acquisition Device," or LAD. The set of rules it uses to find the correct grammar from among the limitless possible grammars consistent with the stimulus he called "Universal Grammar," or UG. UG is, in essence, a set of constraints, guiding the learning system to a small subset of the many possible outcomes it might reach in the absence of such constraints.

I am not going to defend the specifics of Chomsky's proposal here, though later I'll return to it and ask what biological sense it might make, if any, to propose such a device, whether we conceptualize it in Chomsky's terms or some other terms, such as evolved Bayesian priors (Griffiths & Kalish, 2007; Pinker, 1979). Instead, my point is to use the example of learning a grammar from specific utterances to ask what we might mean by a "constraint." How might constraints help a child learn the correct grammar?

One possibility is that they might act like a "restraining bolt." In the film *Star Wars*, the robot R2D2 was fitted with a restraining bolt that kept him from wandering off into undesired places. Like the locked door mentioned above, this constrained him to a small area of space, and when Luke Skywalker unwisely removed it, R2D2 was able to wander off, unconstrained, in a random direction. In the case of language acquisition, an unconstrained LAD would listen to the collection of English sentences in the stimulus set, and would then start using its general-purpose inference algorithms to start generating possible grammars consistent with the stimulus. It would keep generating these, and because there is a limitless number of possible grammars consistent with the stimulus, it would never stop. Like R2D2 without his restraining bolt, it would wander off into the desert of possible but nonexistent grammars, unsupervised.

What role could a restraining bolt play in such a system? One possibility is that it could provide a stopping rule (Simon, 1979; Gigerenzer et al., 1999). The system could start generating possible hypothetical grammars, and when it hit the right one, the restraining bolt could say, "that's it, you got it, you can stop now."

Computer scientists and others who design such systems would tell you that this kind of restraining-bolt system is rather unlikely, at least as a general design for finding specific solutions in massively multidimensional spaces. The reason is that if the only role the bolt plays is to stop a search, there is no guarantee of how much time

it will take the system, randomly running through possibilities, to hit the grammar where the bolt says "stop." It could, in principle, take forever. In massively multidimensional search spaces—which, according to some linguists, the set of all possible grammars is (Evans & Levinson, 2009)—the space of possible solutions expands exponentially as the number of dimensions increases, a phenomenon sometimes known as combinatorial explosion (Cook, 1982).

What is needed, in such cases, is an enabler—something that *makes* the LAD generate the appropriate grammar for English. A mere restraining bolt that waits for a more general-purpose system to hit upon the answer by accident and then says "that's it" is not practical. Far more likely is a scenario in which the LAD makes it such that the correct English grammar, or some small set of variants around it, is the only one that *occurs* to the system. A stopping rule can be used to decide among the candidates, but only when the set is relatively small (Gigerenzer et al., 1999; Holland et al., 1986). Whatever the LAD is, it has to be something that generates the right answer in a positive way, rather than just playing the negative role of saying "no" as wrong answers come along. In other words, it needs to make inductive bets about the (opaque) grammatical properties it's trying to learn, based on the current evidence. One way of modeling such a system is as a Bayesian hypothesis-generation system—a possibility to which we'll return below.

Another classic scenario for illustrating opaque learning situations was introduced by the philosopher W. V. O. Quine. Imagine a visitor in a foreign land where the visitor does not speak the local language. The visitor sees a rabbit hop by. A native speaker of the local language points in the direction of the rabbit and says the word "gavagai." If the visitor attempts to infer what the word "gavagai" means, he is faced with a problem: It could mean "object," "mammal," "ears," "food," "undetached rabbit parts," or a host of other things as well. In fact, the number of possibilities is limitless (Quine, 1960).

In this case, as in Chomsky's, there is a particular kind of problem: how to get the right answer—what the speaker means by the word "gavagai"—given the inputs. The inputs "underdetermine" (do not fully constrain) what the possible intended meaning of the speaker might be, which is why Quine's problem is sometimes called "referential indeterminacy," or the "indeterminacy of translation." However, "gavagai" problems apply to more than just translation. They apply to inductive inference more generally: the inference of anything general from specific cases, such as the meaning of a word from a single use, or general rules from finite instances of their application, as in English sentences (Goodman, 1954; Holland et al., 1986). In fact, as we will see, there is a deep relationship between the induction problems that natural selection solves—retaining designs that bet on the world working certain ways—and the induction problems that organisms solve in order to make everyday inferences,

from deciding what to eat, to what another person means when they say something, to what part of a culturally transmitted behavior to copy. In all of these cases, as in any instance of induction, success depends on certain bets about the world holding true. In the case of adaptations like fins or wings, those bets are engineered into bones and feathers, and in the case of cognitive systems, they are engineered into information-processing rules. And, unlike deductive inference, where a small number of inference rules lead to deductive certainties if the inputs are true, there are no "general" rules of induction, except for a version of the first law: It depends. For example, whatever inference rules might lead to a successful guess about what the word "gavagai" means are not likely to be the same rules that lead to a successful guess about whether a newly encountered mushroom is safe to eat.

There are a host of experimental demonstrations in psychology that show that people solve induction problems all the time, apparently without terrible stress or effort or delay. Even children do it. For example, a variety of studies have looked at how children behave in situations that closely replicate the conditions of Quine's "gavagai" thought experiment. Children will be shown an object and the experimenter will say "Look! A dax!" Then a new set of objects will be brought out and the child will be asked which of them is a dax, thus testing the inductive generalizations they make about the meaning of "dax." On the basis of experiments like this, developmental psychologist Ellen Markman proposed that children make a series of assumptions about the likely meanings of words in a kind of hierarchy (1989, 1990). The "whole object assumption" allows children to guess first that the word refers to a whole object (e.g., rabbit). If they encounter evidence that this isn't true (e.g., they already know the word for a whole rabbit, and it's "rabbit," not "dax"), then they entertain other options, such as that "dax" refers to just a part of the rabbit. The virtue of assumptions is that they generate a possibility right away, rather than searching through a long list of options.

Inductive guesses like this can also be modeled using Bayes' theorem: a rule derived in the 18th century by Thomas Bayes that estimates the probability of a hypothesis being true given the available evidence. We'll cover this in more detail later, but Bayes' rule allows us to model the implicit assumptions, sometimes known as *priors*, that children bring to a learning task—and, as it turns out, it is a very useful tool for modeling inductive bets (Tenenbaum et al., 2006). Whereas Markman's model of children's hypothesis generation is something like an ordered list—first entertain the simplest option, then go to the next option on the list—Bayesian models allow the probabilistic evaluation of multiple hypotheses simultaneously. These can still be "prioritized" in that one hypothesis can be more consistent with the data than others. Both kinds of models have the advantage of actually generating a distribution of possible word meanings based on data—no search of an endless possibility space.

Experimentally, psychologists Fei Xu and Joshua Tenenbaum used a Bayesian model to predict children's guesses about word meanings in a "dax"-type task. As in the "gavagai" scenario, they hypothesized that children entertain different possibilities for the scope of a word meaning, from more general ("mammal") to more specific ("rabbit"), and that the probability of a given meaning was inversely related to the size of the set of things being referred to—essentially, that you're likely to use the most specific word that picks out a given object, but no more specific than necessary. Their model closely matched the inductive guesses of both children and adults, suggesting that some such hypothesis-generation process is going on inside their minds, whether consciously or not (Xu & Tenenbaum, 2007).

As it turns out, we are solving frame problems all the time. There are many other demonstrations showing that problems of indeterminacy are routinely solved as part of everyday life, even by babies (e.g., Bloom, 2000; Gergely et al., 2002; Macnamara, 1982; Tomasello & Barton, 1994). We're solving them every moment we're awake, and we don't even know it. This only began to dawn on psychologists with the advent of the artificial intelligence movement, when they realized how much implicit "knowledge" has to be engineered into robots to allow them to solve even the simplest of human problems, from navigating around a room, to distinguishing apples from oranges, to trying to pick one up. This was where the idea of a frame problem originated, along with the realization that large numbers of such problems exist, and that humans and other animals are solving them all the time (Dennett, 1984).

The purpose of these examples is to drive home this point: Whatever the details of our account of human cognition end up being, they are going to have to entail positive accounts of the structure of the psychological mechanisms that enable the things we see humans, or for that matter any species, doing. To do so, they are going to have to specify how the causal structure of the mechanisms in question, in interaction with the external structure of the world and the internal structure of information in the mind, produce those outcomes—and not just specify the outcomes they rule out. And they are going to need to be able to explain comparative differences between species, such as why humans and chimps exposed to the same spoken languages don't learn the same things about them, or why baboons can't learn to build houses but humans can. We've seen an example of how a positive account of this kind can explain the differences between phototactic and non-phototactic bacteria: The addition of opsins and an internal signaling system fully explains, from a causal perspective, the difference between the biased random walks that phototactic bacteria exhibit in response to sunlight and the non-phototactic behavior of their ancestors. Of course, we have few such complete accounts for any aspect of human cognition, as might be expected for such a causally complex organism (though we do arguably

have good accounts for some parts of human brain systems, like vision). But here's the point: It is this kind of account that we should strive for.

---

By suggesting that all brain mechanisms are specialized in some way, you might think I'm sweeping under the rug the possibility that humans do have some genuinely general-purpose learning mechanisms that can account for important aspects of human cognition in a positive way. To be clear, I'm not denying that many of the mechanisms that psychologists have proposed as "general purpose" might exist, such as mechanisms of conditioning, statistical learning, and the like. All I'm arguing is that if they exist, we should *also* see them as adaptations, engineered by natural selection because of the positive work they do in enabling cognitive abilities.

Let's end this chapter with one more example: imitation. This is a cognitive ability that appears to be quite general purpose, and in a sense you can say it is. But evidence suggests that humans are in many ways better imitators than other species, and it's likely that we possess either new cognitive mechanisms or modified versions of older ones that have added new abilities of imitation that other species don't possess. In particular, imitation illustrates the point that cognition is not simply a matter of passively perceiving information in the world. Perception is important, to be sure, but in order to successfully imitate, you have to make inferences about what it is that you are perceiving (and perception itself is, of course, inferential). Faced with the same information, different species do different things: Some don't imitate at all, and others vary in both how successful they are at it and what they appear to be attempting to do. This suggests that evolutionary events have enabled some species to draw certain kinds of inferences, to entertain certain kinds of thoughts, that other species cannot.

A variety of evidence suggests that humans are sensitive to an aspect of the world to which other animals are not sensitive in the same degree: the thoughts of our fellow animals. This ability, sometimes called mindreading, theory of mind, or intersubjectivity, is extraordinarily complex and will be a central theme of the book, so I will make some initial, perhaps overly simplistic claims that I will later refine. In fact, almost all animals are sensitive *in some way* to the thoughts of other animals. For example, female birds can tell when male birds want to mate, and dogs can tell when another intends to attack. A growing body of research has shown, however, that humans represent mental states of other humans and animals much more subtly than other animals do, and indeed, are able to form *kinds* of representations of mental states that other animals cannot (Baron-Cohen, 1995; Call & Tomasello, 2008; Nichols & Stich, 2003; Saxe, 2006).

Here is an example. Humans can look at another person performing an action—like, for example, attempting to unscrew a lid from a jar, or trying to insert thread into the eye of a needle—and immediately infer what the person is trying to do, or at least make a pretty good guess. Indeed, that is what we generally *try* to do when we see another person systematically engaging in just about any behavior, from peering under their car to jumping up and down on one leg: We ask *why*, or to what end, they are doing it. Although it might not seem so obvious at first glance, there are in fact multiple frame problems being solved every time we do this. For instance, a man attempting to unscrew a sticky lid from a jar is not actually removing the lid from the jar, and yet another person can infer that this is what the man is trying to do, even without seeing the intended outcome. As it turns out, even very young children can make such inferences, as has been demonstrated in an increasingly large and clever experimental literature. For example, psychologist Andrew Meltzoff showed that if children observe a person unsuccessfully attempting to hang a loop of string on a hook, by 18 months of age, they will complete the action and hang the loop on the hook, without having seen the act completed (Meltzoff, 1995). Crucially, this means that the infant is not actually "imitating" the other person, if by "imitating" one means doing exactly what the infant sees the other person doing. Instead, the infant is inferring the other person's *goal* and then bringing that about, in some cases by performing an action that the infant hasn't actually seen yet (and sometimes, skipping steps they have inferred are unnecessary in order to achieve a goal in a different way than they have seen a model performing) (Carpenter et al., 2005; Gergely et al., 2002; Hamlin et al., 2008; Meltzoff, 1995).

In the literature, various technical terms have been attached to different forms of imitation based on what is being imitated. The detailed copying of actions themselves rather than attempting to infer and achieve the goal is sometimes called "blind imitation" or "mimicry," and going straight to an outcome that has been seen without necessarily copying all the steps taken to achieve it is called "emulation" (Want & Harris, 2002; Tomasello & Call, 1997; Whiten et al., 2009). Children can do both of these, and we will see examples later in the book—and the contexts under which they occur, which tell us something about the underlying mechanisms. Here I want to focus on the second, inferential kind of imitation, which I'll call *goal imitation*. What I mean by this is imitation in which we don't slavishly copy every detail of the behavior, but rather infer the means/ends relationship inherent in the other person's actions and attempt to reproduce it.[6] Sometimes, this involves

[6] Goal imitation, as I am using it here, corresponds closely to what is sometimes called emulation, or reproduction of the inferred goal (outcome) but not necessarily the steps to reach it. The reason I'm using a slightly different term here is because I want to allow for the possibility that certain of the steps themselves may be inferred to be part of what is to be imitated (and

copying steps even when we're not sure exactly *how* they produce the outcome—but we're much more likely to copy them if the person we're imitating seems to be doing them on purpose (Lyons et al., 2007). And sometimes, as I mentioned above, we can even jump straight to the inferred goal itself, even when we haven't seen it produced (note here that none of these things is something people *always* do while imitating, because human imitation runs the gamut from being unaware that one is imitating at all, to confused bumbling, to immediately grasping the means/ends relationship; Heyes, 2009).

You might not think that this is a particularly impressive skill, perhaps because it seems so easy for you. If you were to walk by a basketball court and see a man throwing a ball at a column with a hoop attached to it over and over again, what else could you infer but that he was trying to get it in the hoop?

As it turns out, it isn't as simple as you might think. One way we know this, empirically, is because most other animals—even animals that are pretty smart on most accounts, like monkeys—fail to infer goals in potential imitative contexts where humans do (Cheney & Seyfarth, 2007; Visalberghi & Fragaszy, 2002; Fragaszy & Visalberghi, 2004). This is not to say that monkeys are *never* able to infer others' goals and attempt to reproduce them, and there is increasing evidence that monkeys and apes do this much more than we previously thought (Dindo et al., 2008; Whiten et al., 2009). But there are many reports in the literature, both in the wild and in the lab, where other primates fail to make inferences from what another individual is doing that, to a human, would be pretty obvious. Social transmission of skills does occur in primates, and even in some non-primate species like birds (Langen, 1996). But compared to humans, the process appears to be much more slow and painful. It often seems to involve processes like "stimulus enhancement," in which individuals are attracted to the outcomes that others are producing and are thereby inspired to try to figure out how to do it, but they must engage in much more individual trial and error to "get it" than humans would (Whiten et al., 2004).

Consider, for example, social learning of tool use in capuchin monkeys. Capuchins are quite clever foragers who use tools to process foods designed to thwart access, like hard nuts and seeds with painful spines, and who are known to transmit foraging skills socially in the wild (Perry, 2011; Perry & Manson, 2008; Ottoni et al., 2005). A spectacular example of this is nut-cracking by the species *Cebus apella* in Brazil, where the monkeys use specially selected rocks to crack nuts placed on anvil stones

thus could be considered part of the "goal" of the action in the broad sense; e.g., performing an action with a certain style). To the extent that "emulation" refers only to outcomes, it might have a slightly narrower scope than goal imitation as I am envisioning it, but in most cases, the terms could be used synonymously.

while standing bipedally to hurl the rocks at the nuts. Inexperienced individuals seem to realize who is an expert and preferentially watch them, but it takes them at least two years of individual trial-and-error learning to become proficient at the skill. Rather than grabbing stones and attempting to hurl them at nuts when they see others doing this, novice monkeys begin by playing around with the nuts and banging them on things. It takes years—not hours, minutes, or days—to "get" the relationship among hurled stone, nut, and anvil (Ottoni et al., 2005).

For some time, it was thought that capuchins do not truly imitate at all, but rather that they benefit from the increased learning opportunities of having experienced conspecifics do things with tools nearby, without actually *copying* what they saw the other monkeys doing (Fragaszy & Visalberghi, 2004). In support of this, Fragaszy and Visalberghi summarized the results of many laboratory studies showing a seemingly frustrating inability to imitate even the simplest of things on the part of these otherwise smart and dexterous monkeys. In one study, Visalberghi and colleagues took capuchins who had already learned how to obtain food from a "tube trap" apparatus in which a peanut could be obtained by pushing a stick through the tube from one side, and allowed other capuchins to watch them. The key here was that in the middle of the horizontal tube was a "trap," a cup into which the nut would fall if pushed across it. Therefore the peanut needed to be pushed in the direction opposite the cup in order to remove it from the tube. On this task, despite watching experienced monkeys who knew how to get it right every time, naïve monkeys did not learn any faster than they did by individual trial and error (Visalberghi & Limongelli, 1994; Visalberghi & Fragaszy, 2002).

There are many factors that might influence monkeys' performance, or lack thereof, in experiments like this. One factor relates to frame problems. Experimenters sometimes devise tasks like these to test "imitation," as if imitation were a single thing that monkeys either have or don't. But if the task requires multiple inferences or skills, only some of which the monkeys have, they might fail *even if* they could imitate in other contexts. For example, the trickiness of the tube trap depends on gravity and the topology of the apparatus in which the nut is guaranteed to fall in the hole if pushed in a certain way. The inability to "get" this feature of the tube might make learning the task hard, because it might lead you to ignore the relevance of the location of the trap with respect to the nut (of course, you might still be able to learn some kind of rule, e.g., push toward the side opposite the trap, without understanding gravity; but without paying attention to the trap, it might take a long time to entertain this rule as a possible solution to the problem). Interestingly, human children appear to just "get it" without the need for demonstration by about age six, though they are not much better than chimps or capuchins at social learning of the task before that (Horner & Whiten, 2007).

Another factor is what you might call the "naturalness" of the experiment, that is, how much it resembles real-world conditions under which the skill in question, imitation in this case, might be deployed. This is sometimes called the *ecological validity* of the experiment. It might be, for example, that animals in the wild aren't equally able to imitate just any old individual. Perhaps they need to be kin, or more specifically a parent. Perhaps they need to be a friend or a high-status individual in the group, or someone with whom one has had a previous history of trust. In a recent set of studies, Dindo et al. (2008) showed that capuchins can, in fact, learn a two-step procedure for opening a box by watching experts, when those experts are individuals of their social group who have a slightly higher rank and with whom they have a tolerant relationship. Thus, ability to copy might depend importantly on both what is being learned and how—with the scope and nature of the "what" and "how" almost certainly larger for humans than capuchins. Unlike capuchins, for example, you can learn to fly a plane or solve differential equations, and you can learn from reading a book or by listening to a professor you've never met before.

Monkeys and apes can learn via copying, but this does not mean that abilities of social transmission and social learning in humans, apes, and monkeys are the *same*, far from it (Whiten, 2011). While apes raised in human-like environments in which social learning is encouraged can learn remarkable things like referential communication and tool use, their learning process is qualitatively different in many ways, from the complexity of representations they end up forming to their motivations to transmit and use the information. Even if we can teach chimps some symbolic communication, for example, we seem to be in little danger of instigating a *Planet of the Apes* scenario by releasing them into the wild. And human learning involves some odd features like "overimitation," where we copy even irrelevant acts performed by others, as long as they bear evidence of having been done intentionally (Lyons et al., 2007; Whiten et al., 2009). Later I will argue that these are features that have been added on in us: They are the signature of a specialized learning system that exists *because* humans are a culture-dependent species, specially adapted to learn from others. They reflect an underlying bet that when we see someone we wish to copy doing an action on purpose, it's worth copying that action even if we don't understand *why* they are doing it—a fancy form of inference fit to the kinds of culturally structured worlds we inhabit.

The larger point here, and what will be a theme running throughout the book, is that many aspects of cognition that appear simple or that we can describe with a single word, like "learning" or "imitation," might not in fact be single things or skills or mechanisms. They might instead result from multiple interacting mechanisms. Indeed, I'll suggest that most cognitive phenomena we're interested in studying in humans almost always do involve multiple, interacting processes—contrary

to the way "modular" accounts of cognition are typically depicted, namely, as non-interactive. Some of these interacting mechanisms we may share with other species, in either similar or modified forms. These are called *homologies* (Hall, 2003). Others might be special to humans—or, in most cases, involve human-specific modifications to more phylogenetically widespread mechanisms. These are called *derived traits* (if uniquely derived in a taxon, *autapomorphies*). Unfortunately for those who want to isolate the "uniquely human" bits of the human mind, most cases of human cognition probably involve a mix and match of homologous and derived traits. For example, if we consider "the ability to imitate" as a "trait," it seems to be shared by humans, chimps, and capuchins. However, the ability to learn from observation might be boosted by other abilities that humans have, or at least have in some elaborated form, that other primates don't: for example, superior abilities to interpret others' actions as composed of goal-related steps; superior abilities to tell the differences between accidents and mistakes in the behavior we're copying; the ability to form more sophisticated models of the cause-effect relationships in behavior; the ability to form representations of tools as having specific goal-directed functions; the ability to *care* about what others are doing; and more. And to make matters even more difficult, most of these probably involve at least some form of homology (i.e., modified versions of mechanisms that are present in different forms in species that are related to us, a topic to which we'll return in chapters 12 and 13).

---

No matter how you slice it, being able to look at someone else doing something and derive inferences about what they are doing and why is not a skill that comes for free just from having eyes and a brain. It must be enabled, and in cases where species are especially good at it, it is likely to be enabled by multiple specialized mechanisms. Specialized mechanisms are required because goals are opaque in at least two ways (Csibra & Gergely, 2009). First, in many cases of goal imitation that young children solve easily, the goal—in the sense of the goal state, the completed action—is not actually present in the stimulus. Second, the goal—in the sense of the *represented* goal in the actor's head, the thing his brain is trying to achieve—is not observable either.[7] Of course, in cases like observing a skilled nut-cracker, the goal in the first sense

[7] Several scholars have argued for a "direct perception" model of mindreading, in which others' subjective states can be directly perceived in their behavior, without an intervening process of inference (Gallagher, 2007; Zahavi, 2011; see Jacob, 2011, for a critique). Because the notion of "inference" I've developed here is quite broad, potentially covering cases of direct perception, I'll leave aside this debate for now, but note that there is difference of opinion regarding the exact computational steps involved in different forms of mindreading.

is present in the stimulus: The capuchin sees actual successfully cracked nuts. But even in those cases, the mental states of the actor are never observable, including the means/ends knowledge that the nut-cracker deploys in getting the nut open, much of which is in an implicit form that is not even explicitly (consciously) represented by the nut-cracker herself. And yet there is evidence that humans—and in some limited contexts, some of the time, other animals—form representations of mental states that they never observe, like intentions (e.g., to crack a nut, to shoot a basket) and knowledge states (e.g., who knows where food is hidden) (Call & Tomasello, 2008; Gergely & Csibra, 2003; Wellman et al., 2001).

We will go into all this in further detail later, including what it means to form a "representation" of something that one has not observed, like a mental state. For now, what I would like to point out is that whatever cognitive abilities allow human children to rapidly infer a goal from observing an uncompleted action, those are abilities that have been *added on* to the human cognitive repertoire in the evolutionary period of time since the human and capuchin lineages diverged. The interesting question—and the kind of question that I am suggesting in this book that psychology should be dedicated to asking—is the question of the exact nature, the form and function, of the abilities that have been added, the abilities that need to be postulated in order to explain the human/capuchin difference.

How, you might ask, do we know that anything has been added? Couldn't it be that something has been taken away—that, of all the things one might infer from the available evidence, both capuchins and humans can entertain them all, but humans converge on a smaller subset? Couldn't it be that humans are just more observant, focusing more on the right aspects of the stimulus? Perhaps they make inferences like: well, he keeps throwing the ball in that general direction, and the thing in the center of where he's throwing the ball is a hoop—a shape that is complementary to that of the ball—so he must be *trying* to hit the hoop.

I *am* saying that humans make inferences like that. But I'm suggesting that in order to do so, they entertain thoughts that capuchins don't, and quite possibly can't. In particular, to make the inferences humans make, the possibility that the person has the goal of hitting the basket must occur to them—just like the proper grammar of English or the meaning of "gavagai" must occur to a child in order to learn it, and it never occurs to a pigeon hearing the same sentences. Capuchins can be faced with exactly the same evidence, the same stimulus that a person sees, and it never (apparently) occurs to them to think about the relationship between actions and goal. And, as I've pointed out above, the goal itself (both in the sense of having a goal and what the actual goal is) is not in the stimulus, so it must be added, constructed by the mind. There is some inference system, some mechanism that is suggesting the possibility of a goal based on the evidence. Human brains are not merely passively inferring a goal

from what is there; they are looking for one. In other words, the ability to entertain the possibility of goals is an ability that has been added on by natural selection. In the language I will develop further in later chapters, it is a feature of representations—a representational *type* or format—that is not, apparently, available to capuchins.

Goal imitation illustrates a phenomenon that is pervasive in cognition, and that we'll return to throughout the book: the addition of "meaning" by the mind to the raw data of perception. In analogy to the process of colorization whereby color can be added to black-and-white photographs or film, I'll call this *semantic colorization*. What I mean by this is that the mind's inferential procedures add information onto representations of the world that is not actually present in the immediate stimulus: for example, tagging the perceptual representation of an action with a goal, or painting the meaning of what someone is saying to you on top of what are actually just a series of sound waves. In informational terms, this must come in the form of added bits of information, which I'll call *tags*. In phenomenological terms, our brains are built such that we experience this added information as part of what's there in the world: We look at the person straining to open the jar and feel like we perceive the actual *trying*, or hear someone talking and feel as if we're directly perceiving the meaning of what they are saying, as if meaning is a substance flowing from their lips and not something inferred by a very complicated set of neural mechanisms from a series of pressure waves in the air and the visual perception of the person's facial expressions.

Such added features of representations are a dead giveaway for the hand of natural selection, precisely because they are not immediately provided by the stimulus itself. As we will see, these processes of semantic colorization are in fact characteristic of all mental processes, all the way up. They begin at the lowest levels of perception, where even properties of objects like color and three-dimensionality are not directly transported from the world into our minds but are inferred via evolved mechanisms, and continue through to processes of conscious reasoning and decision-making, which traffic in the currency of meanings added by brain mechanisms that evolved to evaluate things in the world based on what they "mean" for fitness. Ultimately, I'll suggest, this is what explains why you are able to look at this set of black squiggly shapes and understand—I hope—the meaning of what I'm trying to tell you. It's a complicated story, of course, and one that involves culture and learning—but also, crucially, brain mechanisms shaped by evolution.

Generative processes like semantic colorization are not easily accounted for in a subtractivist framework. If the mind is an interpretation generator, as I'm suggesting, then this must be because there are mechanisms that evolved specifically *in order* to generate such interpretations, and that use inductive bets to do so. We are not merely passive perceivers, constrained sponges. Instead, processes of cognition are

heavily generative and constructive by nature. And in order to understand the kinds of representations a given species can generate and what it can do with them, we need to think about evolutionary history: the pathway of evolutionary change by which a species' representational capacities accumulated, and how and why each new skill was selected for in terms of fitness benefits and the mind-world fit relationships that delivered them.

In the case of goal imitation, capuchins and humans obviously share much cognitive machinery as a result of their shared primate ancestry, and this accounts for their abilities to perceive objects and actions and to learn via trial and error. But humans are obviously capable of much more sophisticated interpretations of the exact same stimuli. Why? Part of the answer, of course, must involve evolutionary accident or happenstance: In the pathways leading from the human/capuchin common ancestor, our lineage is the one in which certain variants appeared and certain environmental events occurred, and the capuchin lineage is the one in which they didn't. But part of the answer is also undoubtedly natural selection, which, as we'll see, can shape evolutionary history via a kind of self-feeding process in which its own products alter the evolutionary landscape of what's fitness-good and fitness-bad. In the case of humans, part of the answer for why we're so good at imitating is because, compared to capuchins, other humans do lots more things that are worth imitating: We've entered a world in which our own smarts are part of the environment we're adapting to.

# 2

## HILL-CLIMBING

The evolutionary path from bacteria to goal imitation and the many other things that humans can do is a long one. And the journey is not merely a random walk: Setting out in a random direction in possibility space from where bacteria are, you'd almost certainly never get to us. Instead, it's done in a piecemeal way, bit by bit. And as I suggested in the last chapter, many, if not most, of those bits have to have been changes that added something genuinely new. Otherwise, you could take a bacterium, do some whittling, and get a human.

My goal now is to begin pondering the nature of this hill-climbing process when it comes to *thinking*. In fact, there has been a lot written by biologists about hill-climbing in possibility spaces, but much less about the possibility spaces of thought: what it's possible to represent, infer, and learn. So we will need to establish some preliminaries about what it might even mean to ratchet upward in the space of what it is possible to think. This will require uniting three different kinds of possibility spaces—evolutionary, developmental, and informational—and to think about what happens in each of these spaces, simultaneously, when an innovation in information-processing abilities occurs.

In the last chapter, I introduced the idea of a possibility space, the space of all possible states an organism might take. We're going to think of this as expansively as possible to include all the genes the organism might have, all the environments it could possibly inhabit, and all the possible mappings between genes, environments, and phenotypes. And when it comes to all possible phenotypes, this will include things like all possible ways the organism might process information, react to things in the world, and represent the world. In short, all the possible ways it *could* think.

In the 1930s, biologist Sewall Wright introduced the idea of an adaptive landscape to think about how evolutionary processes, including natural selection but also processes such as mutation and genetic drift—random changes in gene frequencies in populations—move populations around in possibility space (Wright, 1932). Wright originally used this metaphor to think about the space of possible genotypes, but it can also be extended to the possibility space of organismic forms, or phenotypes—what

is sometimes called a *morphospace* (Gould, 1991; Gavrilets, 1997; Kauffman & Levin, 1987; McGhee, 2007; Rasskin-Gutman, 2005; Rice, 2002). If you imagine each way in which an organism might vary as a dimension, consider a set of *n* such dimensions, corresponding to the number of ways organisms can vary. The set of all possible phenotypes is an *n*-dimensional space, where *n* is very, very large. Any given organism occupies a point in this space. A population of organisms such as the set of members of the same species will be a cloud, with the degree of scatter representing the amount of phenotypic variation in the population (there are nuances regarding how species are defined, but we will set those aside for the moment). Species that have recently diverged in phenotype space will be clouds that are relatively closer together in this space than more distantly related taxa. However, even different species can occupy overlapping regions of the space in some dimensions, while diverging in others. For example, in the dimension corresponding to limb number, all mammals meet.

Now imagine adding one more dimension: the dimension of fitness, or the ability of a given phenotype to reproduce itself relative to other phenotypes. In Wright's conception, each point in phenotypic possibility space has a fitness associated with it, and we can represent that dimension as altitude. If *n* equals 2, for example—which is of course absurdly small except for the purposes of a thought experiment—then we can visualize Wright's space as a three-dimensional landscape, where phenotype space is arrayed from north to south and east to west as on a map, and fitness equals altitude. Higher-fitness phenotypic variants are uphill from lower-fitness ones, and local optima are peaks. The clouds representing populations will therefore be scattered over this topography—draped across mountainsides or across the floors of valleys—with higher-altitude points leaving more surviving offspring than lower-altitude ones. As Wright pointed out, this difference in altitude corresponds to the strength of natural selection, which will tend to drive populations uphill over evolutionary time, thus the metaphor of hill-climbing (Dawkins, 1997).

This metaphor of adaptive landscapes makes several types of evolutionary phenomena, such as path dependence, easy to visualize. For example, because we've defined our space as the set of all possible phenotypes, the adaptive landscape is continuous, without holes: We can theoretically imagine any phenotype in the space and assign it a potential fitness, were it to exist. But because natural selection only operates on variants that are actually produced, the real landscape is actually *mostly* holes, or open space—or, to think of it a bit differently, populations only actually "visit" tiny parts of the space (Alberch, 1982; Gavrilets, 1997; Rasskin-Gutman, 2005). If they left tracers on the landscape like dots of ink for each point they had visited, the collection of those tracers would look like fuzzy-edged trails, a bit like spray paint, with a thick middle representing the statistical mass of the population's phenotypic distribution and fuzzy edges corresponding to the extremes of phenotypic variance in

the population. Over evolutionary time, these trails would tend to snake uphill—not perfectly, but statistically. And if you examined them at the micro-level, what looked at a distance like a fuzzy sweep of spray paint would actually look more like steel wool, because the offspring of any given individual—lines emanating from the central point represented by a parent—would extend in more or less random directions from that parent. The vast un-inked spaces on the landscape would represent the consequences of path dependence, including, in the case of rugged landscapes with many fitness peaks, the possibility of populations getting "stuck" on local optima, unable to get to other perhaps higher peaks because it would require going down a fitness hill to get there (Kauffman & Levin, 1987).

It's important to remember that evolution is not about change in individuals. Change does happen in individuals, but individuals die. It's about populations. When the cloud of points that represents a species moves over evolutionary time, it is through differential "evaporation" (mortality or failure to reproduce) of some regions of the cloud, and differential growth, or sprouting, of others. And importantly, natural selection is not the only causal determinant of what happens on fitness landscapes, far from it. For one, not all reproduction and mortality is systematically related to organisms' design. Some of it is, and must be, random with respect to design. Remember that the altitude of a given variant on the landscape is the statistically expected fitness of that phenotype, but fitness is not entirely deterministic. Even the best predator in the universe or the most attractive conceivable mate could be crushed by a boulder before reproducing, and even the worst could win the lottery by being the last available mate on a desert island and thereby leaving many offspring. Such random change in the frequency of designs in a population is called *drift*, and it can move population clouds in random directions on the landscape, as well as remove variation from populations over time merely by chance (Kimura, 1983; Ohta, 2002).

Second, the ruggedness of adaptive landscapes can mean that there are whole regions of the landscape—including regions that are fitness-better than the one in which the population currently finds itself—that are never visited, because natural selection will not systematically push populations downhill into troughs that must be crossed to get to other peaks. The details of this, of course, depend on the local topography, and there are no hard and fast rules. For example, some regions of possibility space that appear wide might be easily crossable by certain mutations (Gavrilets, 1997; Kauffman & Levin, 1987). But the point for us is that natural selection is not guaranteed to find fitness optima. Contrary to some depictions of adaptationism, it's perfectly possible to study how designs have been shaped by selection without imagining that they are the best of all possible designs (Breuker et al., 2006).

Finally, because natural selection is the differential survival and reproduction of variants that actually appear, the pathways that evolution can take across the phenotypic landscape depend on the variants that are actually produced by existing developmental systems (Carroll, 2005; Raff, 1996). As I argued in the last chapter, path dependence sometimes has a bad reputation. It's often viewed as being all about constraint—but as we've seen, one man's constraint is another man's enabler. That is to say, if we view the hill-climbing process of evolution by natural selection as the accumulation of design via a ratcheting process, then it is these designs that are the jumping-off points for all future evolution. The region of possibility space where a species currently finds itself may enable it to reach regions of possibility space that simply can't be reached by other species. As we'll see in later chapters, this is likely to be at least part of the explanation for why humans, and not other species of primates living in almost exactly the same environments, entered the uphill evolutionary pathway that led us from the first stone tools a few million years ago to the cities, books, computers, and televisions we have now. This is because some evolutionary changes beget further evolutionary changes by opening up pathways on the adaptive landscape that were not previously available.

Many kinds of positive feedback loops are possible in evolution. Some go under the name of "runaway" selection: If change in a trait alters a species' adaptive landscape in a way that exerts selection on other traits, and if changes in these traits in turn feed back to cause further change in the first trait, a self-feeding evolutionary loop can occur. The best-known examples of this come from the phenomenon of sexual selection, in which the reproductive success of one sex is affected by the behavior and preferences of the other sex. In such cases, changes in one sex, such as preferences for slightly more extravagant ornamentation, can change the adaptive landscape for the other sex: It evolves more elaborate plumage, and this in turn can favor even further changes in the preferences of the first sex, leading to a self-feeding cascade (Andersson & Iwasa, 1996; Kokko et al., 2002).

However, as we'll see, sex isn't the only way that members of a species can either alter the environment or literally *be* the environment to which adaptation occurs, leading to evolutionary feedback loops. In the evolution of cooperation, for example, it is the help that others can provide that impacts fitness. This means that cooperative designs can spread through a self-feeding loop by helping other copies of themselves in the environment (Hamilton, 1964a, 1964b). And in the evolution of mechanisms for cultural transmission, it is the products of these mechanisms—things like languages, artifacts, skills, and the physical environments they shape—that are the environment to which those same mechanisms adapt, in a spiraling cascade (Boyd & Richerson, 1985; Laland et al., 2000).

Of course, in evolution, it is not individuals that adapt, but populations. In these cases, it is populations of *mechanisms* that create positive fitness conditions for copies of themselves, thereby leading the cloud uphill. They are both what structure the environment and make aspects of it "visible" to selection that weren't visible before. For example, the books lying around your house are potentially able to affect your dog's fitness but they don't, because he lacks mechanisms that would allow him to either read books or write them. Baboons *could* learn to make bows and arrows by watching people, but instead they just sit there and stare. Humans can both make and benefit from these things because of a kind of runaway process in which we gradually evolved the ability to both produce such artifacts and learn how to make them by observing others, which requires specialized mechanisms that other species don't possess (Henrich & McElreath, 2003; Richerson & Boyd, 2005). And in fact, every new mechanism makes a new part of the environment "visible" to selection, whether it alters the environment or not. For example, in the case of phototactic bacteria, sunlight was there all along waiting to be used in order to guide movement. It only required the right sunlight-sensitive variant to appear for wavelengths of light to become part of the environment that could lead phototactic organisms up a fitness hill.

What all this means is that we need to think about adaptive landscapes a little differently than real landscapes. We tend to think of landscapes as passive: The shapes of mountains aren't (mostly) altered by our climbing them. Although it's easy to think of organisms and environments as entirely distinct, cases like the evolution of sex, cooperation, and culture show that phenotype space and environment space aren't mutually independent spaces. Our phenotypes are both part of the environment to which adaptation occurs, and constructors of it. The term "niche construction" is sometimes used to describe this reverse form of evolutionary causation (Laland et al., 2000). Evolutionary feedback of this kind means that changes in us—changes in our abilities to communicate, think, transmit information, and alter our environments—have selective consequences on the fitness-goodness and fitness-badness of later design variants. We will return to this in greater depth in chapters 8 and 9.

As I've mentioned, it's impossible to understand evolutionary change without understanding changes in development. Because evolution is the developmental process looped over and over again, evolutionary change *is* developmental change. In this sense, evolutionary space and developmental space are not separate spaces: One is nested within the other. When we speak of the "differential survival and reproduction of designs," we mean that organisms are born, grow, reproduce, and die, and they can in principle affect their own fitness at any point in this process. The design of babies, for example, affects fitness, because if babies don't survive and reproduce,

their fitness is zero. Development is really a progression of phenotypes, beginning in a fertilized zygote and culminating in death, and each of those phenotypes has a fitness associated with it and therefore exists on an adaptive landscape. It is through shaping developmental design, then, that natural selection shapes what organisms look like—or think like—at any given time.

This adds yet more complexity to our evolutionary scenario, because it means that natural selection can shape organismic design as a function of how it *responds* to the environment (Rice, 2002). One way of getting around this complexity has been to gloss over developmental details and assume that there is a one-to-one mapping between genotypes and phenotypes, such that selection acting on phenotypes—which is where it must act, because phenotypes are the vehicles of survival and reproduction—maps directly onto selection on genotypes. This is the approach characteristic of population genetics, which has been an enormously productive approach in evolutionary biology (Grafen, 1984; Maynard Smith, 1978a). Technically, however, we know that it is not really right, at least not in all cases. This is because if we envision developmental space as the space of all possible developmental designs, cases in which a given phenotype corresponds to a given genotype in a one-to-one fashion are but a small subset of all possible designs.

In a manner similar to that of Wright's adaptive landscapes, developmental biologists have envisioned developmental spaces as akin to landscapes. Biologist Conrad Waddington (1957) envisioned the developmental landscape as a metaphorical bumpy surface, with the topography corresponding to the various phenotypes that might be produced during development, and gravity corresponding to the "forces" of development. On the uphill side of the slope are the various environmental conditions that a developing embryo might find itself in, and a ball placed at a point on this uphill edge would trace the phenotypic pathway of development as it rolled downhill, perhaps ending at different phenotypic endpoints depending on where it was released or perhaps not, depending on the topography. I won't dwell on the exact metaphor here because it's confusing to make it fit precisely with adaptive landscapes—among other things, "uphill" and "downhill" mean very different things in the two metaphors. Instead, what is important for our purposes is to envision developmental space as the set of possibilities of how an organism *might* develop, in time, given the organism's genes, its biochemical environment, and its experience of the external environment over time (Goldberg et al., 2007). This is what biologists call a reaction norm: a mapping function between states of the environment, the organism's genotype, and its phenotype (Schlichting & Pigliucci, 1995, 1998; Stearns, 1989; Wagner & Altenberg, 1996; West-Eberhard, 2003). Here, because development is a process of continual change in the phenotype over time, I will also consider reaction norms to have a temporal component, with the state of the phenotype at any given time being

a function of both current and past interactions among environment, genes, and phenotypes, including epigenetic states of the organism (Jablonka & Lamb, 2005).[1]

There are many kinds of reaction norms. Often when people think of the role that genes play in development, they think of what Waddington called canalization: cases in which a given genotype develops into the same phenotype regardless of developmental environment, akin to the downward-rolling ball of development dropping ultimately into the same channel regardless of where it started. We can also call this a flat reaction norm, because it produces the same phenotypic outcome regardless of environmental conditions (at least along the dimension of flatness). It is also what corresponds to at least part of what I'm calling the folk notion or commonsense notion of "innateness" (Ariew, 1999). But in fact, canalized reaction norms are just one possible kind of reaction norm, and perhaps not even the most common one. There are also many kinds of *plastic* reaction norms, in which the same developmental system—including its genes but also its associated epigenetic machinery—will develop differently depending on what environment it's in. We can envision any given reaction norm as having a shape: the shape of the function that maps between environments and phenotypes (think of a graph with environment on one axis and phenotype on the other, and the variety of curves that might map between them). The form that this shape takes depends on the causal stuff of development: genes, of course, but also the properties of all the developmental machinery they build, including the genetic regulatory system and all of the rules for building organisms from zygotes, a complex network of inductive bets instantiated in tiny biochemical machines. The shapes that reaction norms can take are diverse indeed, and can be shaped by natural selection (Rice, 2002). In chapters 6 and 7, we'll return to the question of reaction norms in greater detail, and how their shapes reflect the history of selection that built them.

The final possibility space that we need to consider in order to understand hill-climbing in the space of what it is possible to think is what I will call information space. Here again, we can make an analogy to morphospaces, the space of possible physical forms an organism might take (Gould, 1991; McGhee, 2007; Rasskin-Gutman, 2005). In biology, functional morphology is the field that examines

[1] There are several ways I will be using the concept of reaction norms nontraditionally here (see chapters 6 and 7). For example, traditional conceptions of reaction norms sometimes pick a point in adulthood and call it "the" phenotype or "the" developmental outcome. In contrast, I am treating every point in development as an "outcome," because they are all phenotypes, and all may potentially affect fitness. Thus, I will treat phenotype as something that can vary continuously over the lifetime of an organism, and selection acts to shape the whole trajectory based on its impacts on fitness.

how the morphology of organisms evolves as a function of fit between morphological properties such as the strength and shape of limbs, muscles, teeth, and organs and the ecological contexts in which they operate. In essence, it's the engineering branch of evolutionary biology (Breuker et al., 2006; Lauder, 1990; Vogel, 1988). We can think about mental adaptations in a similar way, as having informational "shapes"—rules, functions, procedures—that determine how they process information and shape the organism's behavior in the world. The study of mind-world fit, then, is a kind of informational blend of evo-devo and functional morphology in which we examine how the computational shape of mental mechanisms influences organism-world interactions, just as the shape of a wing influences a bird's interaction with the air.

In psychology, there have been many metaphors and models of the relationship between the informational shape of the world and the shape of mental mechanisms that exploit or leverage it. Psychologist Egon Brunswik, for example, proposed a "lens" model of cognition in which the organism's judgments about the world are shaped by an intervening cognitive system. He compared this system to a lens that alters the weight given to various cues as a function of how probabilistically predictive they are in the current environment (Brunswik, 1952; Gigerenzer et al., 1991). The shape of the lens, on this view, would be what evolution adjusts. Psychologist Roger Shepard suggested the metaphor of a mirror, in which external properties of the world are reflected in the mind's representations. Although this metaphor is sometimes taken to mean exact reflection of the external world, Shepard recognized that selection warped this mirror as a function of what it reflected and why (Shepard, 1994). Herbert Simon proposed a metaphor of scissors, in which the environment provides one blade, the mind provides the other, and the interaction between the two determines the cognitive outcome: pure interactive causation (Simon, 1990). And psychologist Peter Gärdenfors proposed the idea of conceptual spaces, in which concepts and representations take on a range of possible shapes in a geometric information space (Gärdenfors, 2000). All of these are useful analogies and point to ways that mind-world fit can be formalized in informational terms (Todd & Gigerenzer, 2001).

---

Most of us use the word "information" unproblematically, without pausing to consider what it means. Information is a mathematical abstraction realized in physical entities (Gallistel & King, 2009). A huge field of information science has grown around attempts to formalize the question of what information *is*: the ontology of it, and what it might mean to say there is information in the world in some physical pattern of matter. The problem in this comes from the fact that, intuitively, information is really a kind of psychological concept: To say that I received some information

is to say that I know something I didn't know before. In this regard, the problem of developing a physical, material account of information is continuous with the problem of developing a physical, material account of the mind. Given that what we want to *explain* is the psychological stuff—things like "knowledge" and "beliefs"—we can't have those things be part of the definition of the phenomenon.

This is why information scientists, beginning with the pioneering work of engineer and mathematician Claude Shannon in the 1940s, define information in terms of relationships between physical entities (Shannon & Weaver, 1949). When we say that a dictionary "contains information" about the meanings of words, we are really talking, physically speaking, about patterns of ink on paper. Alone, those patterns do nothing. It is in their *relationship* to something else (e.g., a reader) that they mean something or contain information. The information content of a physical pattern like this in the world has to do with its "aboutness": something that it refers to, even though a dictionary, by itself, doesn't seem to be pointing at anything.

Here, you may detect a deep similarity between the "aboutness" of information and the "for-ness" of adaptations. There are patterns of matter (like ink on a page) that we can say contain information about something, like the meanings of words, even though those patterns are just physically sitting there, don't actually point to any words, and would be meaningless in the absence of a reader. Similarly, there are patterns of matter (like the muscles in a heart) that we say are adaptations *for* something, like pumping blood, even though those patterns are just going through their mechanical motions without knowing what they are for, and would still be adaptations for that function even if they weren't performing it, as in a dead person. Like the patterns of ink, it is something about their causal relationships—what caused them and what they can (potentially) do—that gives them their for-ness or aboutness.

Shannon's definition of information formalizes this notion of "aboutness" in such a way that it can be realized in mechanical machines or modeled abstractly in state spaces. Shannon defined information as a reduction in uncertainty about some possible state of affairs in the world. If I don't know what time it is and then look at my watch, it informs me of the time. I am now much less uncertain about what time it is than I was before. A nice feature of Shannon's theory is that degrees of uncertainty are always relative; there is never exact certainty because, for example, the watch could be wrong, or I could have read it wrong. Shannon developed a formalism that allows us to assign an exact value to information content, to how much information I got from the watch: The amount of information gained is, exactly, the degree to which my uncertainty about the time has been reduced. Notice that this depends importantly on how uncertain I was before: If I just woke up in a dark room and only know that it's sometime between sunset and sunrise, my uncertainty about the time

is reduced far more by looking at the watch than if I just put pasta on the stove a few minutes ago and am checking to see if ten minutes are up.

There is a relationship here to entropy, the continuum between order and disorder that we first saw in the case of Chomsky's theory of language acquisition and later in hill-climbing on fitness landscapes, which Shannon recognized: Information removes entropy (disorder) from a system, taking it to a more bounded and precise area of certainty space than it was in before.[2] We will see this resurface when we start to examine how natural selection pushes cognitive systems uphill in design space. Intuitively, you can see an inkling of it in our phototactic bacterium: The arrangement of opsins on the surface of the bacterium provides it with information about the location of the sun, moving it systematically away from sunlight. In this regard, the light-sensitive bacteria are in a more ordered state with respect to the position of the sun than the non-light-sensitive ones.

What, in the phototactic bacterium, is being "informed"? What part of the bacterium "knows" where the sun is? In any normal intuitive sense in which we'd use the word "know," of course, the answer is none. The bacterium is just being pushed; there is no knowledge.

Nevertheless, in Shannon's sense, information is being transmitted. Shannon regarded his theory as a theory of communication, or inference, between entities in the world: a source of the information and a receiver of it. What makes the relationship one of information transmission is a causal relationship between source and receiver, such that the states of the receiver are in some way systematically tracking—with the possibility of error, of course—the states of the source. I prefer the term "inference" to "communication" here, because communication sometimes implies that the source is *trying* to communicate to the receiver, whereas information can obviously be gained from passive sources like the sun that are not trying to communicate anything (this maps to the distinction between "signal" and "cue" in biology, where "signal" is information transmitted by design, and "cues" can be passively read; Maynard Smith & Harper, 2003). Here, you might argue that the receiver(s)—the bacterial opsins—are not "trying" to infer anything, either, about the source. This, however, is where the element of evolutionary causation comes in: The light-sensitive opsins exist *because* they capture information about the direction of the source of UV light. This is how we can begin to see that talking about the opsins—in particular, how they are arrayed around the bacterium's body—as *representing* the location of the sun could have a physical meaning. Note too that the "signal cascades" triggered

[2] What this means more precisely is that the local entropy of the receiver of information is reduced. The entropy of the entire system, meaning the information receiver plus the world around it, must increase, in accordance with the second law of thermodynamics.

by the opsins are aptly named because they are designed, by natural selection, to communicate information about the direction and intensity of light to the flagella.

What does all this have to do with adaptive landscapes and developmental systems? Where do they meet? When it comes to information-processing systems—structures made of neurons that the developmental systems that build the brain produce—natural selection acts to progressively shape how those systems transform information.

There are many ways to conceptualize what neural information-processing systems do, and lots of contention surrounds the language we use to describe it. For example, there are debates about what a word like "representation," not to mention "computation," might mean in a neural context (Churchland & Sejnowski, 1990; Pinker & Prince, 1988; Van Gelder, 1995). While I recognize the importance of these debates, by and large, I'd like to sidestep them here and remain focused on a Shannon-like view of what information is and does. Most simply, I'd like to adopt an information-processing terminology in which "representation" is meant in the broadest possible sense, as a pattern of neural states that carries information. A representation, in other words, is *about something* in the Shannon sense, although what that is can vary and needn't always be about something in the external world. Neurons can, for example, carry information about some internal state of the organism to another part of the organism, as in the neural system that causes the feeling of hunger. It also needn't be like a little "picture" of what it represents, just as bacterial opsins represent the direction of the sun without forming any kind of picture of it. Similarly, mental (or neural, or cognitive) mechanisms are patterns of neuronal arrangement that *transform* that information in some way.

On this view, one can roughly distinguish between two kinds of things that evolved information-processing systems do: They generate representations, and they operate on them. What natural selection does is to shape the mechanisms that generate representations—of the world, and of inferences that can be derived from the world—and the mechanisms that take those representations, derive new things from them, and use them for the purposes of action. In order to understand how and why thinking evolved, we'll need to think about both of these things: the representational *formats* or *types* that evolved systems generate, and the information-processing *mechanisms* that operate on those representations. This is because, in the course of climbing fitness hills, natural selection can and does produce genuinely new ways of thinking, as we saw in the case of goal imitation: new ways of representing things and making inferences about them. These new ways of thinking take organisms to areas of information space that they could not access before, allowing them to use information that might have already been present in the environment, but that they could not previously use.

---

To make these abstract ideas more concrete, let's return to the bottom of our hill: bacterial phototaxis. Based on the view of information I've just presented, we can see that in the transition from non-light sensitivity to light sensitivity, our bacteria climb not only a fitness hill, but an informational one as well. We can see how this change is instantiated at a mechanistic level: The appearance of a signaling pathway from opsins to flagella renders individuals less likely to die from dehydration or cellular damage from UV light. At the same time, this change allows the organism to make a discrimination, a distinction that it was previously unable to make: It responds to an environmental cue, sunlight, and responds in an adaptive, fitness-good way.

You might argue, perhaps, that this is a weak example of "cognition"—that it's not nearly as complicated as neurons or a brain representing or carrying information about the direction of the sun. Fair enough. Let's move on to brains.

What do neural systems do that causes them to evolve, to push their bearers up fitness hills? They transmit information from one place in the organism to another. Physically, this starts with an event that activates the neuron. For example, the opsins in a rod or cone cell in a vertebrate eye (yes, opsins again) are excited by light, and the cell emits a chemical signal that is detected in the neuron's dendrite (input end), travels as a wave of biochemical events down the length of the cell (axon), and discharges another chemical signal at the output end of the neuron, or axon terminal, which is then either detected by another neuron or, potentially, a muscle cell or some other structure designed to respond to it. Clearly, the reason these things evolved is because they transmit information.

We can easily imagine a very simple case of a phototactic organism where the light-sensitive behavior is mediated by neurons. Light-sensitive cells (an eyespot in a very simple organism, a complex eye in humans) trigger a neuron or set of neurons to fire, which then stimulate muscle cells, moving the organism away, rather like an electrical circuit (Kuenzi & Carew, 1991). Beyond just transmitting information from place to place, neurons can compute: They can add and subtract inputs from different sources, act as "gates" (requiring multiple inputs to reach a threshold before firing), and compute many other functions besides (Koch, 1999; Gallistel & King, 2009). In the case of phototaxis, for example, a certain number of eyespot cells might need to fire at once in order to cause the organism to move, or eyespot impulses could be canceled out by signals from eyespots on the other side, thereby preventing motion.

Consider the changes that happen as a neural system evolves. Imagine a light-sensitive bank of cells arranged as a disc on the surface of a worm. As an initial evolutionary state, imagine that the worm's nervous system is arranged so that it moves away if *any* of those cells fire. Then imagine an evolutionary innovation

such that the worm moves away with a speed proportional to the number of cells in the eyespot firing, or perhaps some even more complex mapping function. For such a transition to happen, the worm needs a little more sophisticated wiring: a gate-type system where multiple signals are summed. And for this to occur, there needs to be a change in the worm's developmental system. In simple animals like the nematode *C. elegans*, we know that the developmental system is highly canalized in Waddington's sense. The effects of genes on development are such that a nearly identical wiring pattern between each of the worm's 302 neurons develops in every single animal, regardless of developmental conditions (or at least under some broad envelope of "normal" conditions—we will return to this sticky issue of what counts as "normal" later) (White et al., 1986). In technical terms, this developmental system exhibits almost no plasticity, no ability to change the developing phenotype in response to conditions. Another way of saying this is that its reaction norm is flat.

In such a simple system, where the wiring pattern is set up so that it works well—the worm moves away from bright light, for example, or moves toward weak light—it's easy to see that most mutations that cause a change in the wiring pattern will tend to mess up the system. Only ones that improve, in fitness terms, the relationship between cues in the world and how the worm responds to them will tend to be retained and spread in the population. For example, a better strategy for dealing with light might appear in the form of a slight change in the wiring pattern. If this variation has fitness benefits over the other variants in the population, it will stick (increase in frequency). The worm's sensory-nervous-motor system can be regarded as a simple computational system: There are inputs (states of the world, e.g., light), computations performed (the pattern of signal summation and subtraction effected by the neural wiring pattern), and outputs: behavioral "decisions" or actual movements on the part of the worm.

It's interesting to consider, here, how design innovations appear. In general, future adaptations begin as "mistakes" in the sense that nothing in the system foresees the future possible use of the change (though they need not and usually won't be purely random in the sense that changes within highly evolved systems needn't go in all directions in phenotype space with equal probability). Indeed, because most changes are deleterious, evolved systems are designed to avoid them in general (Dawkins, 1976). When they are retained, on the other hand, it tends to be because they are useful. In fact, in order for any variant to begin to increase in frequency non-randomly, it must provide an immediate fitness boost, even if the boost is small and statistical rather than absolute. This is because even rare beneficial variants often disappear because random changes in the frequency of variants—drift—tend to have much larger effects than natural selection when variants are rare. But the initial usefulness of a new variant might not (and probably usually doesn't) reflect the eventual

purposes to which they will be put, farther up the evolutionary road. The first eyespot, for example, probably didn't evolve *because* it stimulated a neural pathway that was also attached to muscles in just the right way to be "ready" for the appearance of the eyespot. Similarly, in bacteria, the entire signaling pathway between opsins and flagella presumably didn't evolve all at once. Instead, some initial change in light sensitivity had an effect on the organism—making it twitch, for example, or even just making available a chemical signal of the presence of sunlight that was initially used for some purpose other than motion—which would then set the stage for the appearance of *another* mechanism that coupled that signal to movement, improved the fidelity of transmission between the light signal and the neuron, and so on. This is what the cumulative or additive nature of evolutionary path dependence means: One cell becomes light-sensitive, and this *enables* another system of cells to evolve that can exploit that light sensitivity and use it to transmit information about light elsewhere in the organism. Future changes in the system depend, in turn, on the fact that this channel of information from the outside world has been opened up.

This, in a nutshell, is what hill-climbing is in the world of information processing. A design change makes information available, internally, that was not available before. This in turn changes the shape of the fitness landscape—or, if you will, opens up new possible pathways on it that were not reachable from the organism's prior design state. This change allows for the appearance of new variations to make use of the new innovation, and so on, all the way uphill.

---

You might say that what's going on in the case of the worm is essentially just a reflex: a direct neural pathway between a sensory organ and a muscle, without much if any intervening "representation." The computations are simple at best: not much different than a thermostat, which simply flips a switch when the temperature of a room goes above or below a certain threshold. Moreover, this computational system is purely hard-wired (i.e., canalized, or innate). Not a great model for the kinds of human brain processes I promised I'd talk about, such as thinking. Bear with me, I'll get to it. But simple systems like this do, I think, serve as a good first step for understanding what I mean by hill-climbing in cognitive systems.

Biologists studying the evolution of the eye have realized that successive innovations in eye design—and in the design of neural systems that exploit the information eyes produce—were driven by the ability of eyes to make finer and finer-grained distinctions based on the light hitting them (Goldsmith, 1990; Land & Fernald, 1992; Land & Nilsson, 2002; Lamb et al., 2007). For example, beginning with flat patches of light-sensitive cells on the skin, eyes started to fold inward, creating a curved

cup. That this was a fitness improvement, at least under some conditions, is attested by the fact that this change happened separately in multiple evolutionary lineages including the one leading to us, an example of what is known as convergent evolution.[3] Cup eyes, if they have a small-enough aperture, can be directionally sensitive: Like the seats of a stadium that are illuminated by the sun at different times of day, intensity of illumination on different sides of the cup can be used to tell which direction sunlight is coming from (Land & Fernald, 1992). Of course, this requires increasingly sophisticated neural wiring patterns to make use of this information. Adding discriminative powers leads to more fine-grained behavioral response to light, which leads up fitness hills. And somewhere, far uphill on one of these pathways, are the lens eyes that we currently have, which are called "image-resolving" eyes because they have design features that allow them to resolve an image on the retina (that there are other ways things could have gone is evidenced by the existence of many kinds of eyes other than ours). This image is "detected" in color (three colors, anyway) and transmitted along the optic nerve, a very complicated bundle of neurons, to the rest of the brain.

I've put the word "detected" here in quotes for a reason. A lens eye doesn't detect an image in the same sense that an opsin in a phototactic bacterium detects light. The opsin transmits the message, in essence, "there's light on me right now," and of course the individual *cells* of the retina are doing that, but it would be odd to say that the main function of the eye is to say "there's an image on me right now." Instead, the reason an image-resolving eye evolved is because it allows the brain to make inferences about what the image on the eye *is*, or rather, is *of*: what psychologists would call the *content* of the image. This is, needless to say, a bit more complex than a simple binary discrimination task (light/no light). In fact, the machinery that has evolved to make inferences about the light patterns landing on the retina is mind-bogglingly complex. Come to think of it, you could say that everything we can do with visual information is part of this machinery and is evolutionarily "downstream" (or, in Wrightian terms, "uphill") from the appearance of the eye.

[3] There is a debate (you might be noticing a pattern here) about the degree of convergence versus homology in the evolution of eyes. Salvini-Plawen and Mayr (1977) originally proposed that eyes originated over 40 separate times in distinct animal lineages. However, it has long been clear that there is substantial conservation in the basic building materials of eyes, such as opsins, which are thought to all have a common evolutionary origin. Recently, Gehring and Ikeo (1999) have argued that because a single, highly conserved regulatory gene, *Pax-6*, controls eye development across many species—including species with different eye types (lens, compound)—diverse kinds of eyes are likely to have a common evolutionary origin. Regardless of the degree of homology, there is no doubt that eyes have evolved into many diverse forms, each with a distinct adaptive landscape (Fernald, 2000).

Let us pause a moment to consider some aspects of what's going on with an image-resolving eye. First, note that when I use the word "image," I'm using an abstraction. What there is, physically, is a stream of photons bouncing off some objects in the world (even the idea of an image "of" something is an abstraction, which I'll return to momentarily). This stream of photons passes through the material of the lens, which diffracts it, and then lands, as a distributed field of light energy, on a huge array of rod and cone cells. No single cell in the retina is actually seeing an image. We could say it is "distributed" across the field of retinal cells, but even then, saying "it" is odd because each cell is actually experiencing a photon stream, like a hose of water. At best, the idea of an image is an abstraction in a higher-level state space. We could say the image is "latent" in the stream of photons and that the brain is "trying" to extract it, or "designed" to extract it, but only if we remain aware of what we're talking about physically.[4]

When we say that there are patterns of information in light that the eye is designed to extract, we are invoking a concept that is important in evolutionary psychology: the concept of an *adaptive problem* (Tooby & Cosmides, 1992). One could say that what the neural machinery attached to the eye is designed to do is to solve the problem of extracting an image from the pattern of stimulation on the retinal cells, or, more distally, to solve the problem of figuring out what the image is "of." Indeed, that is exactly the kind of language that many vision scientists would use to describe what the visual system is for (Palmer, 1999). But notice (as many critics of the adaptive problem concept have pointed out) that there's nothing inherent in the light that's causing "a problem": It's just being reflected from objects, passively, minding its own business. The problem, rather, seems to lie in some kind of interaction between states of the world (patterns of light) and the uses that an organism *might* make of them. An organism that *can* solve the problem of resolving an image might get a fitness advantage, but for ones that can't, like phototactic bacteria, there's not really a problem. This suggests that "problems" are not latent in the world per se, but instead are possible pathways on a fitness landscape. They are potentially available to be solved by some organisms whose evolving lineages discover those pathways, but not to those that don't.

In this sense, moving along a path of evolutionary change opens up new problems that were not present before. For example, when the first reptiles began down

[4] There are in fact many ways in which the patterns of information on the retina, and the resulting representations in the brain, are not "isomorphic" to the world. For example, the image of the world projected onto the retina is upside-down, and information from cells that are adjacent on the retina needn't be represented by cells that are adjacent in the visual cortex. Many aspects of representations might stand in a one-to-one relationship to things in the world, but not in a topologically continuous or picture-like way (Palmer, 1999).

the evolutionary pathway of flight, the adaptive problems of flight became visible to selection: Variations affecting flight-worthiness now mattered for fitness, whereas previously they didn't. Similarly, when humans first began to use spoken language, adaptive problems appeared that weren't there for our pre-linguistic ancestors. This is reminiscent of ecological psychologist J. J. Gibson's notion of "affordances": properties of the environment that matter for an organism's functioning within it, but that depend on interaction with the organism's design (Gibson, 1979; Miller & Todd, 1995). For example, an ant and a horse might live in the "same" environment, but the affordances of a blade of grass—and the adaptive problems it poses—are much different for the horse and the ant.

Adaptive problems, then, are not only ever changing, but multifaceted. The blade of grass, for example, poses problems for the horse's visual system in detecting it and for its digestive system in breaking it down. Even *within* vision, in fact, there are many adaptive problems nested within the seemingly "single" problem of seeing things. Resolving an image is one thing, but as I mentioned above, there must be a point to doing so: The point is making inferences about the thing that the image is of. And there are many possible inferences one can make. Different species of organism make very different inferences based on the exact same kind of visual input. Some of this is mere detection: For example, flowers exhibit patterns in ultraviolet light that are available for us to see, but we can't, because we lack the ability to see within the UV spectrum, while bees can. But some involves more complicated forms of pattern recognition and inference, including not just what the visual system can resolve but also specialized forms of learning and interpreting patterns and meaning via experience. For example, you can read this book while a chimp staring at the same page could not, even if he could "see" exactly the same thing as you can.

Vision scientists recognize that the problems of vision are hierarchical. The visual system carves up visual processing into different tasks, with some coming before others in the processing and some coming later. Later processes operate on the results of earlier processes, producing ever-more complex abilities of visual discrimination, categorization, and inference (Palmer, 1999; Ullman, 2007; Yuille & Kersten, 2006). In humans, the visual system tends to first carve the world into objects and then make inferences about those objects based on the representations of them that earlier visual processing has produced. This process of carving the world into objects (i.e., inferring what objects are out there) is sometimes called *parsing*. Parsing, a term originally taken from linguistics, refers to taking something continuous, such as a continuous stream of sound or the continuum of information that enters the visual system, and carving it into meaningful units (the term "parsing" is based on the Latin word *pars*, also the origin of the word "part").

The ability to carve the world into its constituent objects might seem trivial—aren't objects just *there*?—but as I mentioned above, what's hitting retinal cells is like a stream of photons, and there are no objects in a stream. Instead, while objects certainly exist in the world, their properties (such as size, shape, and edges) must be *inferred* from the retinal input. And that is one kind of adaptive problem, or rather, a host of adaptive problems, as vision scientists have discovered.

I won't go into all of the details here—a shelf of books would not be enough to exhaust them—but in a nutshell, the initial stages of visual processing do things like inferring edges in the world from patterns of light on the retina, and these are then assembled into objects via a pattern of inference, since one never sees all of the boundaries of an object at the same time. Even something as seemingly simple as edge detection is in fact not simple at all. Biologists David Hubel and Torsten Wiesel received the Nobel Prize for their work on the area of the brain that detects and interprets the edges of objects, demonstrating types of cells that respond differentially to particular orientations of edges (Hubel & Wiesel, 1968). A large body of research has examined how these cells develop to become specialized edge detectors—or more properly, edge inferrers—and suggests that they may execute a complex type of mathematical function called a Fourier analysis (De Valois & De Valois, 1988). There are many other object- parsing rules and principles as well, such as those studied by the Gestalt psychologists—principles like common motion, which is how our perceptual systems infer that the feet of someone walking behind a fence are attached to the head that is visible above it (Spelke, 1990). And other procedures reconstruct the three-dimensional nature of the world—infer it—from the dual two-dimensional representations delivered by the retinas (Marr, 1982). It suffices to say that the systems that take retinal input and deliver a percept of the world as composed of distinct, three-dimensional objects are complex and specialized indeed.

This sliced-and-diced information is assembled into object representations, which are sometimes called object files (Carey & Xu, 2001; Kahneman et al., 1992; Scholl, 2001). These files are not merely little "pictures" of the object, though they do contain some picture-like, or iconic, information. They also contain sets of pointers, or tags, that attach properties (size, shape, color, location) to object representations and link them to things in the world (Pylyshyn, 2001). Our brains add these pointers because the function of object files is to track certain things about objects, like *which* object it is ("this" cup on the counter instead of "that" cup, even though they both look the same) and *where* it is (if I turn around and can't see either cup, I still maintain a representation of the two, which I can use to reach for *my* cup and not yours when I turn back) (Carey & Xu, 2001).

From a fitness point of view, the advantage of carving the world into objects is so that you can keep track of the objects and their fitness-useful properties in order to

do things with them: object = hamburger; hamburger = food; food = good; reach for hamburger, place in mouth. And as it turns out, even the simplest of object parsing and tracking abilities are not features of representational systems that come for free. For example, human infants at birth have not yet developed the ability to track the location of objects that they aren't currently seeing (Carey & Xu, 2001). What appears to us as a seamless whole—we see an object's shape, know what and where it is, track it when it's out of sight, and so on—is in fact assembled from many different bits of information-processing machinery. Each of these does something special and climbs a little informational hill that the others don't. And this isn't just by chance: A bunch of neurons attached to a retina would not automatically do Fourier analyses, infer the presence of objects, their boundaries, and so on from the retinal flow. They need to be selected to do that.

---

Let us pause to take stock of what has happened in our various state spaces with the advent of the apparently simple ability to parse and track objects.

First, hills have been climbed. Note that it is not necessary for an organism to parse objects to make use of visual information. A visual creature could be designed to move toward or away from objects without parsing them *as* objects. Think of our phototactic bacterium: It responds to light and does so adaptively, but it never does anything as fine-grained as resolving an image of the sun. Slightly more complexly, an animal equipped with a cup eye or even a lens eye could have an algorithm that could detect a greater amount of darkness or light on one side of its cup and move toward or away from it—for example, moving away when darkness begins to rapidly accumulate in a region of the eyespot, perhaps indicating an oncoming predator. "Move away from looming darkness" might work just fine for some organisms, and could well be farther up a fitness hill from other possible designs. In this case, organisms would be responding *to* objects, and in the Shannon sense of representation, they could certainly be said to represent their presence and direction. But an organism that can do this is still not *parsing* objects, computing their boundaries, distinguishing between two that are touching each other, and so on. There is a mind-world relationship in such a creature, a kind of mapping or tracking relationship, but it is not nearly as fine-grained as what we can do.

How is our object parsing different? Look at your sofa and you will see that it is something much more than a dark blob; you will see actual cushions, their contours, their three-dimensional nooks and crannies, where they begin and where they end. If you switch the locations of cushions, you can track where the individual cushions started and ended up, even though the final blob of the sofa is pretty much the same.

And if your chocolate Labrador retriever comes and lies down on your brown couch, you will definitely not treat him as part of the blob.

There are distinct information-processing mechanisms here, each handling distinct adaptive problems. What exactly defines a "mechanism" is a thorny question to which we will return later, but broadly speaking, it is anything that is responsible for a systematic cause-effect relationship in cognition. It is anything that takes inputs (e.g., information from a retinal array), makes some computations, and produces an output (e.g., a carving of "dark blob" into "dog" and "couch"). As I mentioned, vision scientists have discovered that there are distinct mechanisms that do things like find the edges of the pillows on the couch, compute the three-dimensional nooks and crannies of the pillows from both two-dimensional retinal arrays, track the individual identities of the pillows based on the paths that they take when they are moved, continue to represent the stain on the sofa even when the dog is sitting on it, use joint motion (moving together) to represent the dog's tongue and the rest of his head as being attached even though they both move, and so on. These clearly have different functions and presumably different underlying neural forms that cause them to compute as they do. What is frequently not considered, however—or perhaps, taken for granted—is just *how* these mechanisms improve fitness for the organism. One of the central arguments of this book is that this is a step crucial to a proper understanding of mind-world fit. It is not just a luxury, but instead can lead to new discoveries about design. In the case of objects, it's easy to think something like, "well, being able to tell two objects apart is obviously better than *not* being able to tell them apart, so. . . ." And indeed I agree that it *can* be better, though it is not always fitness-beneficial to do so, and not all organisms can distinguish between all possible objects. In cases where we can, the questions are *why* and *how*? What are the specific informational benefits—in terms of the pathway between information in the world and the animal's behavior—that drive the system uphill?

Let us take just one example of the several abilities I've discussed with respect to objects: the phenomenon that psychologists refer to as object permanence. This occurs when you know your cup is still on the counter when you turn around and indeed exactly *where* it is on the counter, or that there is a stain on the couch even though the dog is currently sitting on it (Baillargeon, 1986; Baillargeon et al., 1985; Piaget, 1954).

Just as it was possible to have organisms that can react to things in the world without ever parsing them *as* objects, so it is possible to have organisms that react to objects, and perhaps even parse them as objects, without ever tracking their permanence. What I mean here is forming an enduring representation of the object that lasts after the object is out of sight (it doesn't need to be "permanent" in the sense of lasting forever).

How is it possible to parse and track objects without tracking their permanence? Well, you could imagine a visual system that carves up and makes use of the *current* visual array without much caring about what happens to things in the array when they disappear from sight, or even whether they are the *same* objects when they come back into sight. Tagging something as, for example, "*that* rabbit" and registering it as the same rabbit when you see it again actually requires a separate computational gizmo (or, in all likelihood, more than one), and you could do okay without it. You could have a decision rule that says something like "attack rabbit when you see it" without much worrying about keeping track of particular rabbits. In fact, you could have a pretty sophisticated categorization of objects into rabbits, fat rabbits, and tasty rabbits without having object permanence. And there seem to be some animals that are like this: For example, snakes use cues delivered via multiple senses to locate and attack prey but don't seem to represent that the prey is there when those cues disappear (Gärdenfors, 1996).[5]

So what would be the advantage of object permanence? There are actually several possible advantages, which we can think of as distinct adaptive problems (remember again that "problem" and "advantage" are just two sides of the same coin). One is the advantage of knowing that something is still there even though you can't see it or otherwise detect it. A predator who continued to search for a prey it had seen even after the prey had gone behind a rock or hidden itself in a hole might do better, in fitness terms, than a predator who gave up as soon as the prey disappeared from sight.

Another, separate kind of benefit could accrue from tracking individual *identity* over separate appearances and disappearances. This would require a slightly different kind of computational system than required to just keep looking when an object disappeared. Why bother to track individual identity? Here again, there could be many reasons. To continue with the predation example, the erratic zig-zagging behavior of herd animals like gazelles as they flee a predator might represent a strategy to prevent the predator from targeting any one individual, resulting in the predator shifting from individual to individual in the herd and eventually tiring out (Humphries & Driver, 1970; Krause & Ruxton, 2002). If true, this would select for an ability in the predator to "lock on" to a specific prey individual, pursuing it exclusively until the prey is taken down and not becoming confused by other, nearly

[5] Gärdenfors (1996) makes an interesting distinction between "cued" and "detached" representations, where cued representations require the stimulus producing them to be perceptually present, and detached representations are those that persist once the stimulus that produced them is gone. We will see a similar distinction surface in chapter 5 when we discuss one component of "mindreading" or "theory of mind:" the ability to represent other peoples' beliefs, even when the information we used to infer those beliefs is no longer present.

identical-looking prey crossing its path and momentarily blocking its view—a form of identity tracking.

In fact, there are many other reasons to track individual things over time. These include all of the reasons why we might not want to interchange two individual items (including people) even though they might be difficult or impossible to discriminate along certain perceptual dimensions. For example, my coffee cup might look the same as yours and contain the same kind of fluid, but perhaps I don't want to drink out of your cup because you have a cold. Or my cell phone might look the same as yours when we put them both on the table, but I'd prefer to pick up mine, not yours, when I leave to go home.

So how do human brains computationally solve object-permanence problems? One can imagine a variety of possible ways, but it turns out there are at least two ways our minds actually do it, and they exhibit form-function fit; they use different cues and computational rules, and solve different adaptive problems.

One way of tracking object permanence—and the way that develops earliest in human infancy though not actually present at birth—is spatiotemporal tracking (Carey & Xu, 2001). Spatiotemporal tracking makes use of the fact that physical objects carve out an uninterrupted path in space and time. Objects do not flip in and out of existence—to get from point A to point B, they must follow some uninterrupted trajectory between the two. And when an object is busy occupying some region of space, other objects can't be. Beginning early in infancy, our visual systems make use of this fact, "filling in" the continuous spatiotemporal trajectory of an object, or inferring it, in the parts that can't be seen.

How do we know this? Developmental psychologists have developed clever ways of demonstrating how babies' brains make inferences by, in essence, tricking their representational systems. These methods change the immediate world so that the baby's inferred representations turn out to be wrong, and then the psychologists check to see if the baby is surprised. This method is sometimes called a "violation of expectation" paradigm. Using such a technique, developmental psychologist Renée Baillargeon showed that six-month-old infants assume that a rolling truck cannot pass through a solid barrier that they've previously seen placed behind a screen. If the truck rolls behind the screen on the left and doesn't appear on the right, they aren't surprised, but if it does appear, they are (Baillargeon, 1986). Similarly Karen Wynn has shown that five-month-olds keep track of objects being sequentially added or removed to an area hidden from view behind a screen, keeping count of small numbers of objects in this manner even when only one object was seen added or removed at a time (Wynn, 1992).

Another way of tracking object permanence that is distinct from spatiotemporal tracking is property tracking. Let's say my coffee cup is black and yours isn't. If this is

the only black cup in the room, then I don't need to worry much about spatiotemporal properties—I just look and if it's black, it's mine. Sometimes developmental psychologists call this "kind property" tracking, by which they mean tracking objects via properties that can distinguish kinds of things in the world (Carey, 2009). To test when children can use kind properties to track objects, developmental psychologists Fei Xu and Susan Carey put a toy duck behind a barrier while children watched, then picked up the barrier and observed whether the children were surprised when there was a toy ball there instead of a duck. Initial results found that before 12 months of age, children are not surprised by the duck-ball switch; they know there is *an* object there, because they have spatiotemporal tracking (which develops by at least five months) and saw an object go behind the barrier, but they don't track the objects separately based on other properties (Xu et al., 1999). As it turns out, the story is more complicated, because there is evidence that the development of children's ability to track objects based on a property like shape or color depends on more than just whether it's a "kind" property (Needham & Baillargeon, 2000; Xu & Carey, 2000). It depends, among other things, on what kind of kind it is, what infants know about those kinds, and how they categorize them. For example, Xu (2002) found that verbally labeling kinds with words like "duck" enabled property tracking by 10 months. Luca Bonatti and colleagues found that humans and inanimate objects are tracked separately by 10 months, and infants are surprised when one transforms into the other behind a screen (Bonatti et al., 2002). An even more recent study showed that demonstrating that two objects have different functions (one makes light, one makes sound) is sufficient for 10-month-olds to distinguish them as two separate objects in Xu and Carey's task (Futó et al., 2010). In fact, as we'll see in greater detail below, one could (and adults do) use virtually *any* properties to track object identity, not just "kind" properties like duck properties and ball properties but individual properties, like the scratch that's on my cell phone and not yours, even though they are the same model (and "kind"), or your face versus someone else's, even though you are both people.

---

Why are there two different object-permanence systems, and why might they develop along different temporal trajectories? Evidence of the kind reviewed above suggests that they solve separate adaptive problems, and in order to do so, they use different kinds of information—the information that best enables them to solve the problem in question (note that in a larger sense, both systems are using "properties" of objects; it's a question of the kinds of properties they use). Let's think about the ecological context in which each kind of mechanism is useful—what each allows the organism to do in its environment. A spatiotemporal tracking system allows an organism

to go back and find something even when it lost visual contact with the thing. For example, a squirrel burying a nut, when equipped with this kind of object permanence, can go back and dig for the nut and find it, if nobody has moved it (squirrels and food-caching birds appear to have memory of this kind; Balda & Kamil, 1992; Smith & Reichman, 1984). Notice that this system depends on things staying in their place or following the path they were taking when you last saw them. Notice also that it doesn't "care" too much about the differences between things; if they are switched, as in the duck and the ball, it won't notice.

There are, in fact, cases where animals with such systems are systematically tricked. A good example is the phenomenon known as nest parasitism, or brood parasitism. Some birds, such as cowbirds and cuckoos, parasitize other species' nests by laying their eggs in the nests of other species, alongside the host's own eggs (they wait until the mother is absent to do this). In the case of cuckoos, when the parasite chicks hatch, they push the other eggs out of the nest. Surprisingly, the host mother often cares for these eggs and even the resulting chicks—which are, after all, of a different species than her own, and often many times larger—without acting as if anything unusual has happened (Rothstein, 1990). In some cases, the only cue that the host birds seem to use to track the identities of their individual offspring is that they are whatever was in the nest when they went off to look for food. These parents are like the 10-month-old babies who didn't notice a ball being switched for a duck. The parents' betting strategy has a particular logic, betting that whatever is in the nest is offspring, because in most cases over evolutionary time (as long as nest parasites were rare or nonexistent) that was true.

This is similar to the phenomenon of imprinting in baby geese studied by ethologist Konrad Lorenz (Hess, 1964; Lorenz, 1970). Baby geese, when born, "imprint" on the first object that they see, provided it satisfies certain basic characteristics, like being a certain size and moving a certain way. In Lorenz's case, this included Lorenz himself (or even more bizarrely, his boots). The babies treat this object as their mom, following it around, expecting it to feed them, and so on. Such a mechanism works fine if the first object that baby geese see that satisfies the minimal criteria of their "mom detector" is, usually, mom. The evolution of more fine-grained discrimination abilities will depend on the relative cost of such mistakes, where cost equals the individual fitness cost of the event times its probability of occurring, aggregated over evolutionary time and space. For example, in the case of nest parasites, once they start to reliably appear in a host species' environment, strong selection arises for the host to be able to discriminate between their eggs and parasitic eggs based on properties other than spatiotemporal ones—like, for example, color and size. Given enough time in environments with nest parasites, such discrimination abilities do begin to evolve (Rothstein, 1990).

What about property tracking? The nest parasite example provides a good illustration of the kinds of adaptive problems that property tracking solves—and what it

can't solve. Whereas the pure spatiotemporal tracker is fooled by a switch between two items, the property tracker is not. This is provided, of course, that it can distinguish between the properties of the two items, which is not necessarily trivial at all. In fact, the adaptive hills one can climb when distinguishing between objects are huge and complex, and form the subject of the next chapter. But one can imagine a range from relatively simple property discrimination (e.g., based on color) to relatively complex (e.g., discriminating between two individual faces, or between two individual penguins). The advantage of property tracking is that it doesn't require you to trace the spatiotemporal paths of objects. You can simply walk into a room and, based on properties, say "ah, here's my coffee mug" or "here's my child." The disadvantage—and what spatiotemporal tracking *does* allow you to do—is that it doesn't tell you where to look. You might be able to recognize your coffee mug when you see it, but if you don't have spatiotemporal tracking, you lose all information about where the coffee mug was as soon as you turn your back. The two systems, the two mechanisms, do different things: They solve different problems, they use different information to do so, and they provide organisms with different kinds of fitness benefits.

This illustrates the important point that functional specialization usually involves tradeoffs. Jack of all trades, master of none: A mechanism that is good at one thing is not necessarily good at other things. Psychologists David Sherry and Daniel Schacter (1987) call this the principle of "functional incompatibility": The design features that make a system good at carrying out some function—its principles of form-function fit—might not work well for other functions. They suggest that this explains why humans appear to have one memory system that handles "semantic" memories or memories of facts ("dolphins are mammals") and a different memory system that handles "episodic" memories, in which you can visualize where and when something happened (e.g., when you first saw a live dolphin up close), sometimes called "mental time travel" (Suddendorf & Corballis, 2007). A similar principle might explain why the mechanisms that track spatiotemporal properties of objects are not the same mechanisms that track properties such as what they are or what you can do with them. Indeed, consistent with this view, there is evidence for two separate pathways of visual processing in the brain into which the outputs of the early stages of visual perception are fed: one that handles information about *where* an object is, called the dorsal stream, and another that handles information about *what* an object is and what you can do with it, called the ventral stream (Ungerleider & Haxby, 1994).[6]

So why does one system of tracking object permanence *develop* earlier than the other? Why do babies develop spatiotemporal tracking first and then property

[6] Note that although there is substantial evidence for so-called dual streams in the brain, dorsal and ventral, there remains controversy over whether the "what/where" distinction is the

tracking? We don't yet (and may never) know the answer, but it could have to do with the relative difficulty of building each kind of system: how much experience is required for developmental systems to build the necessary neural machinery, and the relative fitness costs and benefits of having each appear at a particular point in childhood. It might be that once object parsing and a few other things are in place, it's possible to build a spatiotemporal tracking system without needing the infants to have any other property discrimination abilities. Property discrimination represents a very large hill to climb, both evolutionarily and developmentally—and even though infants can begin to do it soon, and adult humans become extraordinarily good at it, there is no reason to build a system that doesn't require it right off the bat.

A second potential explanation has to do with the reasons for which infants interact with objects, and the costs and benefits of tracking individual objects over time based on their individual properties. Consider, for example, a squirrel hiding seeds. It's definitely fitness-useful for him to be able to track the location and persistence of the seeds he's hidden over time. If not, there would be no point in hiding them. But does it help for him to be able to tell the difference between individual seeds, and to notice if one gets switched for another? Probably not. There *could* be a system that distinguished individual seeds based on size, shape, or scratch patterns, but the evolutionary benefits of building such a system would probably not outweigh its costs. Caching animals can remember the locations of hundreds of seeds; imagine the memory required to individually recognize each one (they could, of course, distinguish between seeds based on quality before caching them). Humans, on the other hand, typically *do* remember the faces of many distinct individuals, hundreds if not thousands, presumably because of the fitness benefits of doing so (Bahrick et al., 1975).

In the case of infants and objects, it could be that very young children's interactions with the world are something like this: Here's this truck, and when it rolls under the couch, I want to be able to reach for it and get it back. But it's not until later, when things like individual ownership of objects appears, that kids need to track a *particular* toy or utensil or food item beyond where it is during the course of the current interaction. If so, one might expect development of property tracking to be different for different kinds of kinds and kinds of properties, depending on the fitness benefits of tracking them, consistent with findings on the domain specificity of tracking other humans by Bonatti et al. (2002). Even for adults, it turns out, some properties matter

most accurate way to characterize the functional division of labor they achieve. Other ways of framing the distinction between these pathways include perception versus action and cognitive versus motor systems (Bridgman et al., 1981; Decety & Grèzes, 1999; Goodale & Milner, 1992; Hickok & Poeppel, 2004; Mishkin et al., 1983).

a lot more than others: Adults are much more likely to notice when a living thing changes location in a scene than even a very large nonliving object, like a building, probably because attentional systems are tuned to monitor living things for changes in location, but not inanimate things (New et al., 2007a).

---

What have these examples shown us? First, the progressive addition of finer and finer-grained visual discrimination skills is an example of gradual accumulation of design via hill-climbing. Initially, systems evolve that only crudely register spatial information. But this opens up new pathways on the adaptive landscape: Variations that allow finer-grained spatial discrimination can appear and be retained, leading to the evolution of image-resolving eyes. This, in turn, makes possible the evolution of systems that parse visual input into objects. This paves the way for the evolution of systems that evaluate what those objects *are*, track them, make decisions with respect to them, and so on. And of course, there is no foresight in this process. When cup eyes evolved, nothing foresaw that it was a step toward a system capable of object permanence. Instead, each small step in design space opens up new pathways on the adaptive landscape, pathways that were not accessible before. This is a simple consequence of the fact that evolution works by modifying what's already there. Where you can get to next on a landscape is determined by where you currently are.

In fact, we will see that this path dependence occurs on at least three timescales: evolutionary, developmental, and what I'll variously call the neural, cognitive, or "real time" information-processing timescale. Just as the evolution of object permanence depends on the prior evolution of image-resolving eyes, the real-time spatiotemporal tracking of objects depends on the prior causal step of object parsing. And as we'll see, parallel kinds of causal dependency occur in development too, in which the wiring up of developing systems can depend on earlier-developing systems already being in place.

We've also seen demonstrations of several principles of adaptive specialization. There is the idea of domain specificity: Each mechanism does something a little different with information, and uses different *kinds* of information, to carry out its job. This is another way of saying that each mechanism has an informational "shape," or rather, that each contributes its own little bit to the overall informational profile of the organism in information space. Specialization, in turn, involves tradeoffs: Every evolutionary change has plusses and minuses. By using spatiotemporal information to track objects, for example, I'm not using something else; and I'm opening myself up to deception, in the case of the nest parasites. This doesn't mean, though, that there is necessarily an opposition between specialization and flexibility. Being able to

track objects when they disappear from sight, for example, arguably makes an organism *more* flexible than those that are prisoners of immediate perception. The former can do everything the latter can and more. Flexibility must be enabled. Removing specialization, therefore, doesn't get you more of it.

These examples also show that the nature of adaptive specializations is statistical, in several senses. Natural selection, because it sums up the effects of a design change over many individuals and situations in a population over evolutionary time, in effect "does the math" of costs and benefits, and an innovation will only spread if its fitness benefits outweigh its costs. Adaptations do not need to be perfect in order to evolve; they just need to increase fitness more than other design variants that have appeared. And the inductive bets that adaptations make are statistical as well, in that they won't *always* be right, just often enough to have been selected for. Your object-permanence mechanisms, for example, might prompt you to look for your keys where you last left them, expecting them to be there—but they can be wrong, because somebody might have moved them. And such mistakes needn't be just because of "evolutionary disequilibrium," or evolutionary "lag," in which environments change faster than adaptations can evolve to catch up (Deaner & Nunn, 1999; Laland & Brown, 2006; Symons, 1990; Tooby & Cosmides, 1990). Inductive bets can and will sometimes be wrong even in the environments where they evolved. Indeed, virtually none will be perfectly predictive, as the key example shows. Instead, they evolve and persist because overall, statistically, they increase fitness.

This is worth pausing on, because evolutionary psychology is sometimes depicted as being primarily *about* evolutionary disequilibrium, or being stuck in the past (Buller, 2005; Kanazawa, 2004; Laland & Brown, 2006). Perhaps this is because cases of evolutionary disequilibrium are very useful for illustrating the concept of an EEA, and the fact that adaptation is not a momentary process but one that occurs over long stretches of time leading up to the present. Cases of disequilibrium demonstrate this fact, because while both "adapted to now" and "adapted to EEA" accounts predict the same thing when current and ancestral environments match, only EEA accounts can explain evolutionary lag. An example that's frequently mentioned in humans is our preference for salty and fatty foods, which were rare and valuable sources of nutrients in ancestral environments, but common to the point of harm now. We've seen the example of nest parasites, in which the inductive bets made by the offspring-recognition mechanisms of the parents—feed anything in the nest—can be wrong once nest parasites enter the environment. Similarly, there are cases of disequilibrium that result from extinction of a predator, pollinator, or seed disperser, leaving a species whose design makes no sense except in light of ancestral species that no longer exist. Although it can be difficult to tell if a trait is truly in disequilibrium without measuring fitness and rates of evolution, some intriguing

possible cases have been proposed. For example, biologist John Byers has proposed that American pronghorn antelope, which no longer face strong predation threats, possess traits like fast running speeds that are adaptations to "the ghosts of predators past" (Byers, 1997). And biologist Daniel Janzen has proposed that some rainforest trees in Central America have fruit that evolved to be dispersed by large mammals that are now extinct, such as gomphotheres. This threatens the survival of these trees today, because their seeds need to be passed through the digestive tract of a large animal in order to germinate (Janzen & Martin, 1982).

The appeal of these examples, however, sometimes leads to the mistaken impression that the only explanation for "maladaptive" behavior, or fitness mistakes, is disequilibrium. Even in a state of perfect evolutionary equilibrium, when an adaptation can't be improved—because no variants that would further enhance fitness exist—adaptations can and will sometimes lead to maladaptive behavior. This can occur even when the adaptation in question is operating just as it was designed to do. For example, even a highly adapted prey species will sometimes make a wrong turn and get caught by a predator, and even the most highly adapted predators will sometimes fail to catch their prey, or be fooled into chasing things that aren't prey (Barlow, 1953; Jeschke & Tollrian, 2005; Semple & McComb, 1996). Even well-tuned adaptations sometimes fail; no strategy is perfect. In fact, as (honest) investment bankers will tell you, all strategies are guaranteed to fail under some conditions, and in biology, many adaptive strategies result *mostly* in failure. For example, of the billions of sperm a typical human produces, only the tiniest fraction actually result in offspring; the probability of success, per sperm, is as close to zero as you like. Nevertheless, massive overproduction of sperm is a strategy that has been selected for. Adaptations only need to work better than the alternative variants available in order to evolve.

Taking this into account may matter for our understanding of many psychological adaptations in humans. For example, humans probably have a variety of adaptations that embody strategies for living in groups. One of these might be a concern for reputation: We worry about what others think of us. Presumably, such a propensity to worry evolved because it was better for our fitness than a non-worrying strategy, but it can sometimes lead to mistakes. For example, if you've done something bad, there's no point in worrying about it if nobody can ever find out. Nevertheless, people sometimes do worry, and they even confess to crimes when there is no other evidence against them. Arguably, things like giving to charity under conditions of perfect anonymity may be similar kinds of fitness mistakes. You might think that the only way to explain such "mistakes" is in terms of evolutionary lag, but that's not necessarily the case. An adaptation like adjusting one's behavior based on what others *might* think can be favored in current environments even when it sometimes makes mistakes, as long as its fitness benefits outweigh its costs statistically in the long run. We'll see

other examples of such social biases that sometimes lead to maladaptive behavior, even though they evolved because of their net fitness beneficiality.

---

The details of when and why organisms bet right or wrong, be they bets about food, mates, or reputation, are clearly matters of the first law: They depend on the design of the underlying adaptations and how those adaptations interact with the world. However, it's important to realize that *all* adaptations and the strategies they embody should be viewed in a statistical, probabilistic light. Adaptations are not (usually) perfect little machines, identical in every individual, and operating identically in every circumstance. Every time an adaptation interacts with the environment, that interaction is in some ways unique and different from other, prior interactions. While these facts might seem to bode poorly for a mechanistic understanding of adaptations, they can be conceptualized using the tools of probability theory. While probabilistic models might sometimes be seen as at odds with the evolutionary psychological approach I am proposing here, I think just the opposite is, or should be, true (Bjorklund et al., 2007).

Bayesian models, which I mentioned above, are well suited for modeling inductive bets in probabilistic environments (Tenenbaum et al., 2006). Bayes' theorem provides an optimal policy for estimating the probability of a hypothesis being true (sometimes called the posterior probability of the hypothesis), given some piece of evidence for that hypothesis. Stated simply, this probability equals the probability of observing said evidence if the hypothesis is actually true (called the likelihood), times the probability that the hypothesis is true independent of the evidence (called the prior probability of the hypothesis), divided by the overall probability of observing the evidence. Thus, Bayes' theorem can be written as $p(\mathrm{h}|\mathrm{e}) = p(\mathrm{e}|\mathrm{h})p(\mathrm{h})/p(\mathrm{e})$, where h = hypothesis and e = evidence.

Without worrying too much about the details, the first thing to notice is that this rule is about probabilities: It provides a policy for deriving one probability from other probabilities. Indeed, it works not just for "point" (single) probabilities but for distributions of probabilities over multiple hypotheses simultaneously. The rule is useful because the probabilities on the right side of the equation can often be estimated more easily than the posterior, and in some cases, either estimated very roughly or ignored (e.g., for some applications, $p(\mathrm{e})$ is a scaling constant that can be ignored). It provides a policy for updating one's "beliefs"—representations of the world—on the basis of new data and new experience. And Bayesian models are increasingly being used to model cognitive processes of learning, judgment, and decision-making (Chater & Oaksford, 2008; Tenenbaum et al., 2011). For example, Xu and Tenenbaum's (2007) model of word

learning, which we saw above, used Bayes' theorem to predict children's and adults' inductive guesses about the meanings of words based on a likelihood function that captured intuitions about how broad or narrow a term's referential category might be. The "data," or evidence, came in the form of an experimenter's actual use of the word to label objects in an experimental setting, and when plugged into Bayes' theorem with the likelihood function, yielded good predictions of subjects' inductive bets.

Some evolutionary psychologists might be wary of Bayes. If it's so optimal and domain-general, isn't this just a return to general-purpose cognition, with one mechanism to rule them all? Not if it is interpreted properly. As Bayesian modelers are careful to point out, Bayes' rule is not in itself a neural *mechanism*, but rather a description of an information-processing policy or strategy (Chater & Oaksford, 2008). A Bayesian strategy could be *instantiated* in particular neural mechanisms, bits of neural tissue, but that's a far cry from saying that there is just a single "Bayes box" in the brain through which all information is routed (Gallistel & King, 2009). Indeed, from an evolutionary point of view—not to mention an engineering point of view—it's hard to see how that could be possible.

Just as every mechanism can be thought of as instantiating an inductive bet of some kind, most, if not all, information-processing mechanisms can be modeled *as if* they are evaluating the probability of a hypothesis being true. And they can be modeled as if they have priors, or "prior beliefs," about the world: expectations and commitments that they use to evaluate evidence (cues). Just as our flagellar phototaxis system instantiated bets about the fitness-badness of light in the physical material of proteins and signaling cascades, your visual system delivers a probabilistic bet that object X in your visual field is a cat. At an informational level, this can be described as assigning a probability to a hypothesis (X = cat) based on its observed features (evidence = meowing, cat-shaped) combined with likelihoods (probability of X meowing and being cat-shaped if it is, in fact, a cat) and priors (how likely you are to be seeing a cat in the subway, for example). What physical form these take (e.g., representations and rules stored in some way neurally) is an empirical question (Yuille & Kersten, 2006). But we can, and psychologists do, compare the inductive bets that people actually make with Bayesian models. This allows us to test hypotheses about the specific forms that inductive bets take and their quantitative parameters.

Bayes is just one part of a large landscape of information theory, some of which has already been adopted by evolutionary psychologists. For example, Shannon's theory of information can be described in Bayesian terms, where information gained equals reduction in uncertainty about a hypothesis or a probability distribution of hypotheses (Gallistel & King, 2009). The formal language of probabilities, hypotheses, and information can be used to understand

problems of signal detection (problems of discriminating signals from noise) and the related phenomenon of error management: attempting to minimize the fitness costs of mistakes in "noisy" (random) environments where some mistakes are inevitable (Peterson, 2009; Green & Swets, 1966; McKay & Efferson, 2010; Haselton & Nettle, 2006). Evolutionary psychologists Martie Haselton and David Buss illustrate error-management dilemmas in the realm of dating (Haselton & Buss, 2000). Suppose you're trying to figure out if a potential mate is interested in you. Either he is or he isn't (states of the world), and either you ask him out or you don't (decisions). This creates four possible outcomes: He's interested and you ask him out (a hit, or true positive); he's not interested and you don't ask him out (a true negative); he's not interested and you do ask him out (a false positive, false alarm, or type I error); or he is interested and you fail to ask him out (a false negative, miss, or type II error). In this case, Haselton and Buss conjectured that natural selection might shape men's judgment to overestimate women's interest, given the asymmetries in costs and benefits of the different types of errors. Different systems, of course, will have their own error-management thresholds and biases tuned to the problem at hand and the costs, benefits, and probabilities of different kinds of error.

The probabilistic nature of the world, and of cognition, means that there will *always* be errors. Natural selection can't entirely eliminate them. Instead, it will tune the properties of mechanisms so that they tend to produce the least fitness-costly statistical distribution of errors across the environments where they evolved. Moreover, when we begin to conceptualize evolved mechanisms as making probabilistic bets about the world, we can begin to see that they needn't be as rigid or frozen in the past as is commonly assumed. Rather than mechanisms coming in two types—the narrow, rigid, innate ones frozen in the past and the broad, flexible ones designed to handle novelty—it could be that *all* evolved mechanisms have some capacity to respond to novel situations, and what makes organisms flexible isn't the possession of some all-purpose ingredient, but rather a combination of mechanisms making bets of different kinds. When an organism responds adaptively (or maladaptively) to an evolutionarily novel situation, then, the interesting question becomes: What combination of mechanisms is operating, and how do the inductive bets those mechanisms make, when combined with the world's data, generate the behavior that we see?

PART II

# Information

# 3

# ADDING VALUE

In the previous chapter, I suggested that representational systems progressively add features to representations as they travel through the processing stream of cognition. What lands on the retina is just a rain of photons; what emerges from the retina is a field of neural firings representing the activity of the rod and cone cells. Later, these are parsed into surfaces and shapes, and color and depth are added via a set of inferential mechanisms, each of which adds its particular properties, the outputs of its computations, to the growing representation. And further on, the representations can be built up even more: We can add information about object category (dog, book), individual identity (Fido, *Moby Dick*), and many other kinds of properties as well (growling Fido, the copy of *Moby Dick* that I borrowed from you).

One way to think about this is that our brain is adding little pointers or tags to representations of objects, events, and anything else the mind can represent—like a pointer that says "*my* cup," tracking one of several perceptually identical objects based on its spatial location on the counter, or the tag that identifies an object as a cat (Carey & Xu, 2001; Scholl, 2001; Kahneman et al., 1992). As I've suggested, these are a bit like colorization added to a black-and-white film: The original picture is still there, but additional features are added. There is nothing *on* the cup itself that says "my cup" or on the cat that says "this is a cat." Our brains are painting the scene with meanings that aren't objectively there, semantically colorizing them. The purpose of these added tags is that other brain systems can use them; other systems, for example, can now assume the object is a cat without having to make that calculation themselves.

In a scene from the film *The Terminator*, we're momentarily shown what it looks like to see the world through the Terminator's eyes. As he scans objects, little pop-up menus appear, telling him what the object is and giving him various behavioral options with respect to it. What I want to argue, in this chapter, is that our brains are doing something like that for us all the time. We are a bit like cyborgs walking through a landscape, with little pop-up screens informing us about the things around us: "*This* is good to eat" or "you left your pencil under *this* pile of papers" or

"*this* object coming toward you is a person, and it's George." All of these inferences are added by evolved systems. Without them, "George" would just be a shape on the retina. These systems add value and meaning to what's out there. In fact, they *infer* the value of these things, and not just any value, but their value for *us*, for the purposes that these things can serve in our lives. And ultimately, systems to infer value in this way only evolved because they increased our fitness. We care about the apple because we can eat it; we care about the growling dog because he can bite us; and we care about George because he can be our friend or enemy.

We are moving away from the realm of perception and into the realm of what are sometimes called concepts, thinking, reasoning, and the like: cases in which the meanings of things, rather than just raw perceptual features, come into play. I use terms like "concepts" and "thinking" loosely here because I'm not concerned, for the moment, about the precise distinctions between, for example, "conceptual" and "perceptual." Crucially, all information-processing involves inference and representation ("thinking"), though one could say that the farther representations proceed along in the processing stream, the more conceptual the distinctions our brain makes become.

Importantly, what conceptual systems do—and "conceptual" carvings begin very soon after information enters the brain—is not so different than the functions that the object parsing and tracking systems carry out. Like parsers, conceptual systems sort things into categories or along continuous dimensions, giving us the dogs and the cats, Democrats and Republicans, the nice days and the not-so-nice ones. As we proceed up through the processing chain—although I will argue that it is not, contrary to some conventional wisdom in psychology, necessarily linear—the parsings become more and more fine-grained, and also more "abstract," less derivable from raw perception alone. And, like trackers, conceptual systems put pointers on things that allow us to track them in space, time, and memory, depending on what they mean to us in terms of how we will interact with them in our lives. Like the child retrieving his truck that has rolled underneath the couch, our brains spatiotemporally track George so that we can find him when we need him. But they also tag him with a host of properties, ranging from "George" to "Democrat" to "likes science fiction movies," which allow him to pop up when our brains search for objects with the relevant properties. I will argue that our brains make these conceptual parsings all the way up into the highest levels of abstract thought we can entertain, and that while not every detail of every parsing is specifically due to natural selection, the *types* of parsings that we make are.

But there is something else that our conceptual systems need to do: they need to tell us what attitudes to have toward the parsings they make. How should we react differently to a rock or a dog or a person—are they just different kinds of objects with different sizes and shapes? What's the difference between the Bible and *American*

*Psycho*: two books, one older than the other? And Republicans versus Democrats: two kinds of people on CNN? Without a value system—a principled interface with the systems that drive our muscles and tell us how to act—a categorization scheme is useless. And, as I've suggested, categorization mechanisms only evolve *because* of their effects on our actions, and ultimately, on our fitness. Otherwise, they would be selectively neutral. This is where emotion comes in, motivation and feelings, and the many links between thinking and action: between concepts and what we *do* with them in the world (Bechara et al., 2000; Damasio, 1994; Fessler, 2006; Öhman, 2006; Tooby et al., 2005). Contrary to the commonsense view that separates "cognition" and "emotion," I am going to suggest that the conceptual/thinking parts and the emotional/motivational parts of mental activity, while they may involve distinct psychological mechanisms and processes, cannot be decoupled. They fit together like a lock and key, for a reason: Conceptual distinctions exist only because of their importance for, ultimately, action, the things we can do with them; and emotional and motivational systems exist only because the world is parsed in ways that allow them to make us decide to do the right thing.

---

In philosophy, an ontology is the answer to the question of what exists, and how it should be carved up into categories, or kinds. Scientific disciplines come with ontologies, which entail what are sometimes known as "ontological commitments": certain assumptions about how reality works (Lewis, 1999; Quine, 1960). For a long time, for example, physicists were committed to a basic distinction between matter and energy. These were different things, both real but not the same, with different properties. Now, physicists regard these as merely different versions or arrangements of some fundamental, underlying stuff. The matter/energy distinction is still real, on some levels of description, but the ontology has been expanded to include other descriptive levels. Evolutionary biologists, in turn, have traditionally been committed to a difference between genotypes and phenotypes; these are different because of their causal properties, even though, at some level, they are made of the same stuff (atoms), and even though it's becoming increasingly accepted that genes can and should be considered part of an animal's phenotype (Mahner & Kary, 1997). And the sciences of mind have ontologies too that carve up mental processes into kinds. A typical psychologist's ontology presumes that, for example, memory and reasoning are different kinds of things (Bilder et al., 2009).

Human minds—any minds, in fact—have ontologies as well. The perceptual and conceptual parsings that the mind performs on its representations of reality establish what we can think of as intuitive ontologies (Boyer & Barrett, 2005). Although we

might never contemplate the nature of our ontological commitments, we nevertheless have them, at deep levels of our thinking. They are embedded in the inference systems that give rise to our representations of the world. For example, the fact that object-parsing mechanisms sift input from the retinas, looking for edges and other cues to put objects together, suggests that these mechanisms "expect"—bet—objects to be out there. They expect the world to contain objects that can be detected and parsed by their neural algorithms. Everything that uses their output assumes there are objects too. In a world without objects, these mechanisms might make some major mistakes.

Intuitive ontologies involve not just sorting the world into things, but what you might think of as *principles*. Physicists don't treat matter and energy just as arbitrary kinds, like red and blue M&Ms—they do different things and have different properties. Similarly, we will see that mental mechanisms carve reality into kinds and embody implicit assumptions about how those kinds of things will behave. We already saw one example of this: Spatiotemporal tracking mechanisms "assume" that physical objects don't blip in and out of existence, but rather follow continuous, unbroken paths. Again, this assumption is not written into the mechanism anywhere as an explicit principle. Instead, because the function of the mechanism is to generate representations of where an object is at any given time, the computational system that predicts where the object will be computes its location when out of sight using a function that assumes a continuous spatiotemporal path. This reflects a commitment about the way material objects behave, but it is a commitment engineered into the structure of the mechanisms that generate representations of reality. As we've seen, those commitments can be demonstrated experimentally by causing reality to appear to violate them and seeing whether observers are surprised. Many such commitments of our ontology of physical objects have been demonstrated, including that two material objects cannot occupy the same space at the same time (Baillargeon et al., 1985), that force can be transmitted from one to the other in collisions (Leslie & Keeble, 1987; Michotte, 1963), and that they respond to the force of gravity (Needham & Baillargeon, 1993).

As I mentioned above, conceptual systems do not produce representations just for fun. Intuitive ontologies produce representations of the world that interact with our motivational and action systems, telling us what significance the different kinds of things in the world might have for our behavioral decisions. For example, our minds assume that the world contains objects that are foods, or potential foods, and colorizes our representations of food with valuations that index their potential nutritional value, as well as other kinds of value like social prestige. Foods are kinds of things that can taste good or satisfy a craving; they can be ingested and given to others as social offerings. Other kinds of entities do not necessarily have these properties.

The reason that I have bothered to introduce the idea of an intuitive ontology is not because I want to imply that our mental mechanisms are like internal scientists or philosophers. It is because I want to emphasize an aspect of (much) evolved mental machinery that is frequently not appreciated: that the role of this machinery is as a *framer* of our thoughts. If you comb the literature on evolutionary psychology, you will frequently see reference to things like "innate representations" and "innate content" (Elman et al., 1996). Even if you want to think about representations as "little pictures" in the head—which I've suggested can be misleading if taken too literally—intuitive ontologies do not typically provide the content of the pictures themselves. Instead, they provide the machinery that allows the picture to be painted, and painted in such a way that it can then be used and interpreted by other inference systems, such as decision-making systems. In this sense, intuitive ontologies can be seen as *ways* of interpreting or framing things. Psychologist Frank Keil calls these "modes of construal" (Keil, 1995, 2003). Philosopher Daniel Dennett calls them "stances" (Dennett, 1987). As Keil's experimental studies of thinking have shown, the same object can often be interpreted in different ways for different purposes: We can change our framing of it. This is an important but counterintuitive kind of flexibility, and at odds with the way that many people think of automatic, inflexible "modules." For example, we can think of a person both as an *object* that occupies space—which entails a set of ontological commitments or inductive bets about objects, such as that he will move backward if pushed—but also as a *person*, a social actor, which entails a set of ontological commitments that are applied to people but not rocks. These include the assumptions that they can act in goal-directed ways, have beliefs, belong to social categories like friends and mates and kin, and more.

This, in turn, implies that the evolution of new ontologies does not necessarily preclude the use of evolutionarily older ones. Instead, the addition of new ontologies adds new inferential powers and therefore new kinds of flexibility, new ways of dealing adaptively with the world. For example, if the account of goal imitation I presented is right, an animal that doesn't have an intuitive ontology of mental states may be able to process information that is "about" mental states in some sense, but not in the same way as us. Like us, it can see another individual cracking nuts, which is mentally caused and goal-directed, but doesn't see that action *as* goal-directed. Once it has an inference system for treating this information specifically as mental state information—not just raw pattern detection, but a set of principles for handling the special properties of mental causation—a whole new possibility space of inference and decision-making is opened up.

---

The set of ontologies humans have is large, as large as the number of ways we can frame and interpret the world. We share some of these ontologies with other animals. For example, many assumptions our visual systems make about the properties of solid objects are probably homologous with versions of those same assumptions present in other primates: shared via descent from common ancestors (Van Essen & Gallant, 1994). Others appear to be derived, unique to our lineage or a small set of related lineages. Some ontologies might be specific to humans alone (e.g., an ontology of grammatical language, an ontology of tools made to carry out functions, an ontology of ethnic groups; Hauser et al., 2002; Jackendoff, 2002; Mithen, 1996; Richerson & Boyd, 2001). Some might be specific to primates (e.g., an ontology of friendship and alliances; Cheney & Seyfarth, 2007; Silk, 2007). Some might be broader still (e.g., an ontology of kinship). I've already mentioned the ontology of objects and some of the inductive bets it uses to interpret the world, ranging from bets about the topological continuity of objects, to spatiotemporal continuity, to principles of physical causation such as the transmission of force via collision. Humans also have complex *social* ontologies, ontologies of people, social groups, minds, and behavior, which we'll come to soon. And, as we saw from the example of physics and the matter/energy distinction, people can develop new ontologies—usually, with the help of many other people, in the form of culture—that haven't been specifically selected for by evolution (Carey, 2009). We'll come to that as well. But first, to give a sense of what an ontology means in terms of psychological specialization and mind-world fit, let me examine an example of a specific ontology that humans have—and that we share parts of, but perhaps not all, with other species: the ontology of food.

The intuitive ontology of food illustrates several things about the design of evolved inference systems more generally: in particular, their intimate relationships to emotions, decision-making, and learning. In the case of food, it's easy to see how the whole purpose of the system is that it adds value to things that otherwise are just objects. Food also shows that learning and evolution are not separate explanations for psychological phenotypes, because learning is part of the very design of systems for representing, evaluating, and making decisions about food.

Every animal has an ontology of food. That is to say, every animal carves the world into things that it doesn't eat and things that it does, or things that it could (there are some exceptions, such as sponges, which are animals with no nervous or digestive systems). As they move through the world, animals don't just carve the perceptual array into objects, they carve it into objects and then ask, "What good (or bad) is this thing to me?" In survival terms, it's hard to imagine many objects—except perhaps mates, predators, and a few other things—that have a greater fitness importance for the organism than food. And because animals are mobile and ingest

food through their mouths, they are faced with different adaptive problems when it comes to food than are, for example, plants.

But why do I go so far as to say that every animal has an ontology of food? I say this because even the most simple-minded of animals doesn't just carve the world into food and non-food objects. Objects that are categorized as potential food are in turn treated in special ways; the organism has special *attitudes* with regard to these things that involve emotions, decision-making, motor actions, and bodily functions. When an animal stops and tries to decide "am I going to eat this?," a complicated inferential system is activated, connected to emotions and the body, that is fit to the purposes for which food will be used.

One of the most interesting and well-studied aspects of how animals respond to food is that it involves learning—indeed, very specialized learning. If you're familiar with Pavlov and his dog salivating at the sound of a bell, you might think: What more general-purpose learning system could one imagine than associating a sound with a food? As it turns out, food learning (along with the learning of fear; Öhman et al., 1976) was one of the first major challenges to the "equipotentiality" assumption of the behaviorist school of psychology, which held that any two stimuli are equally easy to associate. In a now-famous series of studies with rats, psychologist John Garcia and colleagues showed that rats more easily learn to associate the taste of a food with a sensation of nausea, experienced hours later, than with a later sensation of electric shock (Garcia & Koelling, 1966). This makes good adaptive sense. The evolved function of nausea is in part to inform the rat that something it just ate is harmful, and this provides an opportunity for the rat to learn—and not just learn about *anything*, but about something it just put in its stomach. Since then, there has been a modest but growing amount of research on the form and function of the food learning system (Cashdan, 1994; Galef, 1993; Hills, 2006; Rozin & Kalat, 1971; Rozin & Fallon, 1987; Shutts et al., 2009).

When I say that there is an ontology of food, then, what I mean is that there is a set of mechanisms that takes some objects and places them into the category of "things that might be eaten" and then treats them differently—submits them to a different series of inferential procedures—than other objects. The category of potential food could potentially start as broad as nearly all objects, but it rapidly gets narrowed down during the development of any organism. Usually we think of foods as "stuffs" or "substances"—meat, cheese, vegetables—but not necessarily so; many organisms, including us sometimes, are predators that look at animate objects and think about them as potential food. From all the things that *could* be food, the brain's task is to decide which ones it should actually put in its mouth and swallow.

To do this, food evaluation systems are intimately connected to systems or processes that we sometimes categorize as emotions, feelings, or drives, like hunger and

disgust (Fessler & Navarrete, 2003; Hills, 2006; Rozin & Fallon, 1987). These, in turn, are clearly linked to memory and decision-making. For example, the Garcia effect, mentioned above, is related to what are called learned food aversions: After being made sick by something, such as your seventh tequila shot at a New Year's Eve party, your brain can semantically colorize that food item for a long time after the event, making the sight, smell, and thought of it repulsive (Midkiff & Bernstein, 1985). The mechanisms of our food learning systems also make use of direct feedback from the physiological systems that monitor the nutrient content of food after it's entered our bloodstream, such as glucose, fats, and other nutrients, leading to positive semantic colorization of nutritious foods. The systems that govern our attitudes and behavior with respect to food are tied to these systems like a thermostat is tied to a thermometer: Just as a thermometer tells the thermostat when to turn on or off, both our long-term food learning and our short-term eating decisions are tied to our stomachs, which tell us what was good and what was bad, which things to put in your mouth and which not to, when to put them in your mouth, and how much (Rozin & Fallon, 1987; Ritter, 2004; Wurtman & Wurtman, 1986; see also Scheibehenne et al., 2010, on the importance of visual cues).

The two sides of the motivational coin with respect to food are hunger and disgust. These emotions, and subtle shades of them, are part of the palette that our brains use to semantically colorize food, adding a pointer or tag that decision-making systems guiding behavior can use ("Don't eat this!" or "This would taste great in your mouth"). There are many striking demonstrations of this in psychology, but perhaps the most well known are by psychologist Paul Rozin and his colleagues who showed, for example, that people are averse to eating a piece of fudge shaped like dog feces, even when they *know* it's fudge (Rozin et al., 1986). This makes sense, given the function of the food evaluation system: to evaluate the food item *before* you eat it, on the basis of cues that can be assessed at a distance, such as appearance and smell. Here, something like associative learning might be a good explanation for what's going on, but it must be a special kind of associative learning, since you don't have to have put dog feces in your mouth to learn to not ingest things with their shape and color.

Another example is the phenomenon known as "pregnancy sickness," in which foods that once seemed delicious now seem (and smell and taste) disgusting, probably for the purpose of protecting the developing fetus from things in foods that could cause birth defects, illness, or death, such as pathogens and protective chemicals made by plants (Fessler, 2002; Profet, 1988). Pregnancy sickness demonstrates perhaps most clearly that the value of things is painted onto them by the brain, and does not reside in the things themselves. Otherwise, how could mothers report that things that were once craved, like coffee or asparagus or sushi, suddenly become so disgusting that they can't bear to even smell them?

Hunger is the opposite side of this coin. Both "disgust" and "hunger" describe not just phenomenology—the way it feels to be disgusted or hungry—but links to decision-making and behavior. Disgust turns off appetite and causes avoidance, making us want to move away from certain objects because the disgust system is betting that they can make us sick. Hunger leads not only to approach but to a desire to ingest food items and even, as Pavlov showed, an anticipatory digestion reaction, showing again the intimate connection of body and mind (Pavlov, 1902).

Hunger, of course, also involves learning—not that we learn hunger itself or how to be hungry, but rather we learn what to be hungry for (Rozin & Kalat, 1971). Here again, there is a mysterious-seeming but fascinating process of associative learning via experience, because we often get cravings for specific items. This is likely to be, at least in part, because our physiological nutrient evaluation systems have experienced particular foods and evaluated their contents in the past, and then updated the representational parts of our brains—perhaps even down to the perceptual level—so that when we are lacking salt, for example, those potato chips look especially good, or that Gatorade is especially thirst-quenching. Although the details are, as far as I know, still not entirely known, it seems likely that natural selection has engineered a communication channel that goes both ways between these various systems. The search for nutrient-specific cravings has turned up some broad categories of cravings, such as those for carbohydrates, salt, and sugar (Denton, 1982; Rodin et al., 1991; Wurtman & Wurtman, 1986). Certain kinds of foods, like meats and sweets, show up high in studies of cravings across cultures—moderated, of course, by culture-specific food traditions (Whitehead, 2000; Zellner et al., 1999)—and some cultures even have words or concepts for specific kinds of hungers, like "meat hunger" (Shostak, 1981). Older studies with rats showed that they preferred foods with specific nutrients they were lacking, like vitamin B, though vitamin-specific hungers remain controversial in humans (Richter et al., 1937). There are also the odd cravings and aversions seen in pregnancy, some of which can be explained as aversions to potential pathogens or substances that would be damaging to the fetus, and some, such as ingestion of soil, may reflect cravings for nutrients lacking in the diet (Fessler, 2002). All of this suggests a complex appetite-regulating system whose logic we have not yet fully deciphered, but that is clearly designed to categorize foods and tag them in ways that make us desire or avoid them.

"Deliciousness updating" goes so far as to influence spatial cognition: how organisms represent their environment spatially, and how they decide to navigate through it. Perhaps it's not a coincidence that your shopping cart feels as if it's pulling itself toward the dessert aisle. There is a field of study in animal behavior called foraging theory, which develops mathematical models of how organisms should decide to allocate their time across the spatial environment as a function of the nutrient contents

of food and how they are distributed in space (Charnov, 1976; Hills, 2006; Stephens & Krebs, 1986). For example, the "marginal value theorem" of optimal foraging theory predicts that a bird eating berries off a bush should stay in that bush only as long as it is getting nutrients at a rate that is at least equal to the average amount of nutrients it would get from starting on another full bush, divided by the time it would take to get to that bush and start eating, on average. Optimal foraging requires specialized learning abilities: Not only must a bird's brain be able to calculate a running average of nutrient intake rates based on current experience, it must also tie these to an internal map of the environment, which contains a representation of, among other things, the average likely distance to the next bush, how many berries it is likely to contain, and so on, tallied over the bird's experience.

While there has been debate over exactly which of the sub-models of foraging theory best predicts animals' behavior—for example, exactly which currencies they are maximizing and minimizing, and whether there is a single model that can apply to all foraging species—there is substantial evidence that animals' decision-making involves a complex calculus of costs and benefits, updated by learning, that includes how they should allocate their time across the landscape of foraging opportunities (Stephens & Krebs, 1986; Perry & Pianka, 1997). Many animals appear to construct internal "maps" of their environments that allow them to navigate to locations like food sources or nests using representations that act as pointers to places in the world (Gallistel, 1990). Humans also appear to maintain internal representations of the value of different food types and how they are distributed over the landscape, though again, the simplest optimality models do not seem sufficient to capture the complexity of human foraging decisions (Hill et al., 1987; Krasnow et al., 2011; New et al., 2007b; Smith, 1983). In part, this may be because the models used in studies of human foraging employ potentially crude measures of "value," such as raw caloric content. In fact, it seems as though the logic of optimal foraging in humans is one in which what counts as "value" is quite flexible, since optimal foraging principles seem to govern things like how car thieves search for cars to steal, how people search for information on the Internet, and how they search their memory for answers to word puzzles (Bernasco, 2009; Hills, 2006; Hutchinson et al., 2008; Pirolli & Card, 1999; Wilke et al., 2009).

In many species, including humans, learning what to eat involves a special kind of learning, which we will see again in future chapters and which is of crucial importance in the design of many human cognitive systems: *social* learning (Bandura, 1977; Zentall & Galef, 1988). Psychologists distinguish broadly between two kinds of learning, individual and social. As social as it might have felt, your bad experience with tequila at the New Year's Eve party was an example of individual learning, because the information your brain used to update its attitude with regard to tequila

was based only on your own individual experience with the tequila and its effects (though your probability and quantity of drinking could have been influenced by the presence of others). The same goes for Garcia's rats who learned to avoid certain foods because of their own individual experiences of nausea. Social learning, on the other hand, involves leveraging *other* individuals' experience in updating one's own representations. It is betting, in this sense, that what other individuals know is fitness-useful. This kind of learning obviously has several advantages—not needing to get sick in order to develop an aversion to something is one—as well as disadvantages, which we will examine in more detail later.

Perhaps the most interesting and well-studied example of social learning about food comes from Norway rats, studied by psychologist Jeff Galef and colleagues (Galef, 1996). It starts in the womb: Fetal rats are sensitive to food particles in their mother's bloodstream that cross the placental barrier. Rat babies after birth are more likely to eat what the mother ate while she was pregnant (Hepper, 1988). This is "social" in a sense, but it also involves, as does all social learning, individual experience: The reason the food particles are in the mother's bloodstream is a social one (they reflect her food choices and not those of the fetus), but the experience of them is individual, that of the fetus alone.

Imagine the representational mapping process that must take place in order for this to occur. Something in the fetal rat's blood system detects food particles, which are then mapped to smell, which is mapped to behavioral decisions after the baby is born. This is a specialized learning system if ever there was one. The advantage of this system, of course, is that the baby benefits from the mother's lifetime of learning about what is good to eat. In fact, the process continues after birth when the baby develops preferences for food detected in its mother's milk—presumably by the same, or a similar, mapping mechanism (Galef & Henderson, 1972).

In addition to being an example of social learning, this is an example of what is sometimes called *prepared learning*: learning caused by a system specialized to make inductive bets about the structure of a particular learning domain (Barrett & Broesch, 2012; Seligman, 1970). In fact, there are many names for specialized kinds of learning, including guided learning and experience-expectant learning (Gould & Marler, 1987; Greenough et al., 1987). However, as I've pointed out, *all* mechanisms have to be specialized for something; every adaptation has a domain. Thus, there is a sense in which all learning is prepared learning, because any learning mechanism, even the most general-purpose one, must make at least some assumptions about the structure of its inputs and how that structure predicts the future (Gallistel, 1990; Tenenbaum et al., 2006). We'll be returning to this issue in more detail in chapters 6 and 7 when we talk about reaction norms, of which learning mechanisms are a part.

Once the rats have grown old enough to forage on their own, there are several different kinds of food learning that continue to shape their developing preferences. One involves following adults around and foraging along with them, but also foraging where they smell adults' scent (Galef & Beck, 1985). Another involves directly smelling the breath of adults: Rats develop a preference for foods that they smell on the breath of others (Galef & Wigmore, 1983). Interestingly, Galef and colleagues have discovered one of the proximal cues that the food valuation system uses to update its representations: carbon disulfide, a chemical present in rats' breath that, when placed on food, causes rats to strongly prefer it (Galef et al., 1988).[1] And, of course, as Garcia showed, the rats have good old-fashioned individual learning, such as learned food aversions for foods that made them sick.

We will return to social learning in later chapters, but let me note a few things right now. First, the social learning strategy is intimately related to (and evolved because of) the rats' ecology, both their foraging ecology and their social ecology. Rats are omnivores, meaning both that it would be difficult for natural selection to "innately specify" the positive and negative values of all the foods they might encounter, and that rats' dietary niche is a risky one; there are plenty of things they might eat that could make them sick or kill them. This is where the rats' social ecology (group living, often with relatives) comes in, because it allows the rats—or rather, their learning mechanisms—to leverage the collective experience stored in others, trusting others' experience when doing so outweighs the risk of trying new foods oneself (Boyd & Richerson, 1988; Galef & Whiskin, 2008). This kind of leveraging is likely to be especially important in humans.

A second thing to note about rats' social learning mechanisms is that they clearly embody a betting strategy. Given that rats sometimes make mistakes, what is in a rat's bloodstream or on a rat's breath is not *always* something good to eat. Galef and colleagues observed that rats will strangely acquire preferences for foods that they smell on the breath of other rats that are sick or dying (Galef et al., 1983). Why does this persist? One possibility is that no better variant of the learning system has appeared for natural selection to favor, such as one that takes into account the health of the rat whose breath is being smelled. This would be a form of evolutionary constraint. Another possibility is that taking into account whether or not another animal is sick is not, in fact, a fitness-increasing strategy on average. Modeling results by Jason Noble and colleagues suggest that in the rats' case, this may be true: If most toxins in the rats' environment are lethal, then encountering a sick individual is a

[1] The effect of carbon disulfide for rats in some ways resembles what monosodium glutamate, or MSG, does for humans. MSG is thought to be a cue to the protein content of foods and colorizes them as delicious. This is now exploited in certain cuisines as a culturally evolved strategy to improve the appeal of food (Krebs, 2009).

very rare event, since most individuals who have tried toxins are dead. Thus, the benefits of a sick-discriminating system might not be outweighed by the costs of maintaining it, including mutation load (Noble et al., 2001). If so, this would be an example of mind-world fit wherein the absence of a finer-grained discrimination system is explained by the frequency structure of the environment: Cases where the finer-grained system would be fitness-good are too rare to support it.

Though we might not care to be reminded of it, humans share some aspects of both rats' foraging ecology (we are omnivores) and their social ecology (we live in friendly social groups). Although there has not to my knowledge been demonstration of exact analogs or homologs of the rats' social learning mechanisms, such as breath-smelling, there is research suggesting a large role for social learning in the human acquisition of food preferences, as well as disgusts. Anthropologist Elizabeth Cashdan, for example, has suggested that children have a specialized food learning system that has a "critical period"—a concept reminiscent of Chomsky's proposal that the language acquisition device has a critical period in which it acquires its settings during childhood (Cashdan, 1994). According to Cashdan, very young children start out being open to eating just about anything, and their dietary preferences are acquired positively—like rats—by narrowing down to the subset of foods that they grow accustomed to eating with their parents and family (see also Shutts et al., 2009, for evidence that food preferences can be socially transmitted). The flip side of this coin is an aversion to unfamiliar foods that increases with age. People grow to find things disgusting that are clearly perfectly edible, since people elsewhere in the world, or maybe even next door, eat them. Acquisition of food taboos may represent an extreme version of this, where some foods become dispreferred in a highly socially marked way, probably for additional social reasons (Fessler & Navarrete, 2003; Whitehead, 2000). Of course, much of food learning is likely to reflect individual learning as well, but clearly much of what we like to eat or don't eat has to do with what others around us eat or don't eat. At least some of this—but probably not all, as the arbitrariness of food taboos suggests—has to do, like rats, with leveraging collective knowledge about what's good and bad in the local environment (food taboos are likely to be at least somewhat different because many perfectly nutritious foods are tabooed). In all likelihood, use of social information is a design feature that has been built into human (and other mammals') food valuation systems because of the fitness value of this social leveraging, and we will see this leveraging of social information in many other cases of human learning as well. The point is that for humans, as for rats, this kind of learning design makes sense, given what human ecology, both foraging and social, has been like over evolutionary history. It exhibits fit not just to the world, but to the way we live in it.

There is one final aspect of our intuitive ontology of food that I'd like to mention. It is one that might be specific to humans or perhaps to certain kinds of carnivores more generally: the intuitive ontology of meat. I mentioned earlier that a curious but functionally important aspect of our intuitive ontologies or modes of construal is our ability to frame shift: the ability to apply different ontologies (at least sometimes) to the same object or situation, as the situation demands (Dennett, 1987; Keil, 1995). In this regard, objects aren't necessarily "just" objects. A cow has physical, object-mechanical properties like a rock: You can push it, it can collide with things, it can't occupy the same space as another object, and so on. It also has mental properties unlike a rock: It can turn if you shout at it, run away from things that it sees, and so on. And as it turns out, it also has substance properties: It is made of stuff like bones, blood, and meat.

In a small but growing literature, psychologists have begun to show that there is an intuitive ontology of substances, or "stuff" (Bloom, 2000; Pinker, 2007; Prasada et al., 2002). Substances are, of course, also objects, like the meat that's in a cow or a blob of yoghurt on your plate; but your brain doesn't necessarily treat them *exactly* the same way as other objects. A pile of rice is made of a bunch of solid bits—objects—yet we can think about it as stuff, because it behaves as a fluid on certain scales, as when pouring rice into a bag to take home. The distinction surfaces in English when we say things like "give me *some* rice," the same way we'd ask for "*some* yoghurt," as opposed to "give me five hundred rices," or more properly, 500 grains of rice (Pinker, 2007). The linguistic distinction is between what are sometimes called count nouns and mass nouns, and a growing body of research shows that children are sensitive to this distinction when learning words (Bloom, 2000; Carey, 2009). For example, Sandeep Prasada and colleagues found that English speakers assume "some dax" refers to stuff, like yoghurt, shaped in random blobs; but when it appears to be carefully shaped so that its "objecthood" matters—like yoghurt shaped as a letter—they are more likely to assume "*the* dax" refers to it (Prasada et al., 2002). Interestingly, Gil Diesendruck and colleagues have shown that these inferences also depend on assumptions about how the shape of the object was created, including whether it was intentional (Diesendruck et al., 2003.)

Some recent unpublished research of my own suggests that our minds treat meat very much like a substance—more like a substance than an object, in fact—but only once it's been removed from the body of an animal or once the animal is dead, not while it's still part of a moving arm or leg. In interviews with children about animals and the meat that comes from them, I found that for food substances like meat, children don't particularly care about the shape the meat takes as much as they care about its substance properties: its consistency, its flavor, its freshness, and other properties that matter for food valuation systems. Moreover,

food substances can be subdivided in any way and still retain the affordances that matter for them as food: half a steak is still a steak. The same is not true for many of the affordances of cows that matter for our interactions with them, such as their ability to move and respond to stimuli. Cut a cow in half and those properties, which psychologists call properties of *intentional agency* (agency for short), disappear. Not surprisingly, then, children agree that half a steak is still a steak, but half a cow is not a cow—a linguistic signature of two distinct ontologies at work, an ontology of substances and an ontology of intentional agents (Pinker, 2007).

This ability to flip frames or modes of construal makes adaptive sense: The different modes of construal we can apply to a thing depend on how we are going to interact with it, for fitness purposes. When meat is still part of the moving cow, and we're hunting it, herding it, or bringing it into the barn for the night, it's more useful for our minds to treat it as part of a whole-object intentional agent (things with thoughts and feelings and goals) and to leave substance inferences for after the hunt is over. Once we get down to butchering it, on the other hand, it's no longer useful to worry about the meat running away and much more useful to think about how to divide it according to edibility criteria. Consistent with this, developmental psychologist Tanya Behne and I have shown that when an animal dies, children turn off attributions of agency properties to it, recategorizing the object from an agent—a living, animate thing—to a substance (Barrett & Behne, 2005). This kind of flexibility in thinking about the same underlying stuff, or object, makes sense in a hunting, meat-eating species like ourselves.

---

From these examples, we can extract a few general points. First, each new ontology that gets added on to the mind is just that: an addition. As the meat/cow example showed, adding a new ontological distinction does not necessarily prevent you from thinking about things using the systems you had before. What it does is add on a new way of thinking about them, framing them, which looks for additional properties and inferences that previous systems might not have highlighted as being important. Having a specialized food valuation system, for example, would not necessarily preclude learning about food through preexisting learning systems that are not specialized for foods. But categorizing food as a special kind of thing and then linking it to learning and inference pathways whose inductive bets are tailored to that category of thing enables the system to learn much more rapidly, targeting what the system already knows, or bets, are the relevant properties.

Second, the addition of each new system, because of the new capacities it adds, pushes the system a little bit up a fitness hill. I mean this in at least two ways. First,

to the extent that the new inference system leads to better decisions, it also leads to higher fitness. This is why, for example, we have a brain that predicts where an object will go if we throw it, or couples visual cues to the nutritive value of food. Second, whatever design features have arisen in the new system have taken it up a *particular* fitness hill in design space, which will, in turn, have a strong effect on the future directions in which the system can evolve.

Finally, these examples demonstrate several aspects of mind-world fit. They show that the design features of representational systems exist because of what we do with them, ultimately culminating in decisions and actions. When we represent physical objects, possible things we could do with those objects are activated: As for the Terminator, fitness-relevant affordances of objects are highlighted for us by our brains. When we represent foods, value is painted on, and in particular, fitness value, as reflected in hunger and disgust. And these design features bear a statistical betting relationship to the adaptive problems they solve. They are right often enough, but not always right. They don't use *all* the information, but the information that has been reliable enough to build a rule for how to use it into the cognitive machinery (Gigerenzer et al., 1999). And finally, these examples demonstrate, over and over again, the first law of adaptationism: The design features of different systems are different, depending on what they evolved to do and the evolutionary trajectory they took to get there.

---

Let's pause for a moment to ask a skeptical question: How do we know evolution has anything to do with any of this? And even if evolution does have something to do with it, how do we know that it isn't something very minimal, like building general-purpose systems such as pattern-learning neural networks that merely happen to be able to do these things without having been selected specifically to do so?

That's possible, of course. To answer the question "What does evolution have to do with it?," we need to ask several other questions. First, for any given process, we need to ask *how* it does what it does, informationally. What are the minimum computational requirements that we need to posit in order to explain what the system does? And second, we need to ask what natural selection needs to have put in place to produce a system with those properties. What is the minimum evolved structure that we need to posit to explain the computations? This will give us the simplest or most parsimonious explanation, which is not necessarily the most likely one. To arrive at that, we must look at all of the available data, including things like cross-species comparisons, developmental studies, and everything else at our disposal, and ask: What is the most *likely* explanation for what we observe, given all the available evidence? It is by asking these questions that we find out what evolution had to do with it.

By "computational requirements," I don't necessarily mean anything fancy or technical. I just mean that we should look at what the system does and ask what it would need to get it done. For example, both rats and humans construct a category of possible food items and narrow it down over their lifetimes. This implies an interface that decides which items in the world go into or out of this category, which in turn implies sensory analyzers and object parsers that can sort and lump foods according to similar or dissimilar properties. It also implies a learning mechanism or mechanisms that narrow the category down. When we look at the nature of the learning processes in rats and humans, we see required design features—only some kinds of learning mechanisms will do the job (Gallistel, 1990). Learned food aversions require a specialized learning mechanism that tags certain items as bad based on individual digestive experience, and this can't be simply a "general" associationist learning system because, as Garcia showed, it knows to tag *foods* as bad things after getting sick, and not other things that might have intervened between the eating event and the nausea event.

Moreover, in this case, just one mechanism won't do. Things like aversions to feces-shaped fudge can't be explained by experience-based learning of the same kind, since we can have such aversions without ever having eaten feces (i.e., it's not a learned food aversion based on eating and then getting sick; there must be some other mechanism, e.g., one specialized for smell and that doesn't require illness per se). And in the case of rats, of course, learning from smelling someone else's breath requires a special detector that uses chemicals in the breath as a cue to tag certain foods as good. All of these, in turn, require some kind of categorization system that carves up foods into categories *within* the overarching "potential food" domain. Otherwise, none of these mechanisms would know how broadly to generalize the good/bad tags. If I smell Spam on another rat's breath, what do I tag as good? Spam alone? Anything with pork in it? All meat? Anything salty? There is a "gavagai" problem in any learning situation. And finally, all of this learning needs to be hooked up appropriately to decision-making systems that govern behavior, systems that don't just generate the thought "oh look, a feces shape," but an actual visceral aversion.

You might posit that all these things—and/or properties of other systems I've mentioned, such as systems for predicting how solid objects will move and react to contact—could be done with general-purpose mechanisms that did not evolve specifically for that task. As I said, it's certainly possible. But you'd have to specify how a general-purpose system would learn to get the job done. For example, in a Bayesian model, you could specify the priors of the system and how its hypothesis space is structured. In all likelihood, you'd need some kind of structured Bayesian model to do the job, because the way food learning and decision-making systems interact isn't easily accountable for merely by altering the probability weights across a single set of

hypotheses through learning. Foraging isn't merely a matter of associating cues with behaviors, since optimal foraging rules instruct the organism how long to stay, when to leave, and what to do in between. This is better approximated by things like if-then rules and logical operations. Bayesian models are appropriate for modeling how such a system is adjusted by experience; they can take on many degrees of complexity. But in the language I'm using here, to the extent that the learning system has priors and structured hypotheses, it is specialized for its learning domain.

You might be in agreement that the examples I've given, such as foraging, are good examples of evolved systems, and that positing specialized formats and rules is plausible and possibly even necessary in these cases. But you might think I've picked particularly good examples to make my case, examples where it's easy to see that evolution is probably involved. For example, many psychologists would argue that what I'm talking about in all these cases are examples of "input analyzers," where it's easy to think of cues that could be systematically detected, and to build a simple system at the level of perception to detect them (Carey, 2009; Fodor, 2000). But some would argue that there are other more conceptual processes, more remotely detached from input, that surely arise out of more general-purpose machinery. For example, much of human thought has to do not just with the perceptual cues of what's being thought about, but with its *meaning*.

I agree. What I am claiming is that evolved systems of the kind I'm describing are what add meaning to perceptual stimuli; they create the meaning dimensions over which other systems compute, such as the systems that cause you to associate the smell of basil (a food) with your mother's kitchen (a special socially marked place or context), or for you to paint a type of meaning onto the visual stimulus of someone cracking nuts—the goal—that a capuchin monkey cannot.

The claim that natural selection can only create perceptual input analyzers is a specific case of one of the misunderstandings that I mentioned at the beginning of the book: that natural selection operates mostly at the level of perception and less so as one proceeds to more abstract, conceptual processing in the mind. The more conceptual and abstract you get, the harder it is to imagine "detectors" for the categories of things that people think about—like "fairness," or "justice," or "friendship"—much less any built-in rules for dealing with them. In fact, in the next chapter, I will suggest that even systems that allow people to entertain very abstract concepts must be enabled by selection. For each *kind* of information our minds can represent and handle, there exist evolved representational systems that make possible those kinds of representations. Our job, therefore, is to come up with an appropriate science of what kinds of information our minds evolved to handle, and which systems are specialized for which kinds of information: What are the joints, or the continua, between the kinds?

Higher-level conceptual systems face the same problems that the object system, substance system, and in fact any information-processing system face: They need to know what they are looking for in the incoming information stream, and how to turn it into the kind of representation they can use. This might involve, in some cases, minimal cue use or representational format change, but in each case, evolution must have made a positive addition to the system to allow it to do so. Such systems *do* have "input criteria"—they involve input analyzers as all computational mechanisms do—but the inputs are abstract indeed, and often not what you'd think of as simply perceptual. Indeed, there is good reason to think that evolution has managed to build adaptations to things that you'd be hard-pressed to see or touch at all: things like thoughts, language, and culture.

# 4

# SOCIAL ONTOLOGY

Social reality is an odd kind of reality. It is full of things that nobody has ever seen or ever could see. We regularly speculate about what so-and-so thinks, if they really believe X, whether they are angry or guilty or jealous, or what someone meant by what they said. Even when words are physically present as sound waves that can be perceived, what we are really trying to do is infer something invisible, hidden, underneath: what the other person *means* or intends to convey. We attribute intentions and motivations and interests to things like companies, political parties, and members of other social categories whose ontological status as "things" is not entirely clear ("tree huggers," "soccer moms"). And there are many, many other invisible properties that our brains treat as real, things that are sometimes called "socially constructed," like the value of a dollar bill or a house, which not only aren't perceivable but arguably aren't even properties of the thing itself at all (Searle, 1995).

And yet, these concepts do not refer to nothing. They are not merely hallucinations (or at least not all of them are, though this is a domain that notoriously lends itself to them). There are real causal processes and patterns underlying all of the phenomena I mentioned above. When we say that someone is "jealous," for example—assuming we're right—we've identified something going on in his brain that does have a causal influence on his behavior and lends some predictability to his future actions. And as we have seen, when the world contains recurring causal patterns that have real-world consequences that can impact fitness—as the mental processes and actions of others certainly can—natural selection can engineer adaptations to exploit or leverage these patterns. A good example from biology is kinship, or genetic relatedness: As the biologist William Hamilton showed, adaptations can evolve that make inductive bets based on the *probability* of shared genes, even when such probabilities are not only impossible to directly observe, but are also statistical in nature (Hamilton, 1964a, 1964b).

Of course, there are many types of leveraging relationships that natural selection could build. For example, the behavior of snakes that makes them dangerous is certainly caused by processes occurring inside their heads, but natural selection could build a system that says, in effect, "stay away from snakes," without any

tracking of the snakes' mental states at all (of course, such a system would still need a way to reliably identify snakes, hook this up appropriately to decision-making and motor systems, and so on). Indeed, natural selection has built such systems in many animals, including humans, who have a short-cut snake reflex of this kind (Öhman et al., 2001). But humans also possess genuine ontologies of the social: systems designed to comb perceptual input for socially relevant information, to parse it, and to add value according to a set of implicit ontological commitments embodied in inference rules. These systems paint the world with social meaning, using categories and concepts that do, in fact, have causal, predictive power, although they are best thought of not as "pictures" of anything—because there is nothing to be a picture of—but rather as cognitive pointers or summaries, representational abstractions that work because of their statistical relationship of fit to the complex causal processes in the world, instantiated across individuals, that they summarize.

What is crucial to realize here is that it is not just that things like "beliefs," "desires," and "preferences," along with "styles," "trends," and "norms," are invisible—it is that there is a sense in which they do not really exist. Much mischief has surrounded this idea, so I want to be precise. Beliefs, desires, and intentions aren't "things" in the sense of being *objects*. If you imagined, for example, that one day it would be possible to perform a brain scan on someone and then point to the resulting picture and say, "There! *There's* the belief!," you'd probably (as far as we know) be wrong. But as the philosopher Daniel Dennett has stressed, there is something causal about organisms, due to the ways they process and store information, that concepts like "belief" and "desire" *do* capture (Dennett, 1987). If I always keep my keys in a certain drawer and look for them there, then a system that represents me as "believing" they are there will nicely predict my future behavior and will capture the fact that I will look in the drawer even when the keys aren't there, if I didn't know they were moved.

Thus, although many psychologists would argue that beliefs and desires are only ever operationalized in actual behavior, our brains nevertheless paint these things onto our representations of people as if they were properties like color, shape, or weight (e.g., Mike likes ice cream; Bill believes in Jesus). This is reminiscent of the way in which owners (and others) represent their houses as having a value that, while changeable, is kind of like a tag attached to the house—even though economists would argue that the "value" of a house is only ever operationalized by what people are actually willing to pay for it. So real-seeming are such tags that some homeowners might believe that even if everyone on earth died, their house would still have a value of a million dollars. In both cases, the shorthand is *factually* wrong on a certain level, and yet is heuristically right because it generates useful predictions: e.g., how Bill

might respond to comments about Jesus, or what kinds of offers will be made when your house is put on the market.

In the social realm, our brains by design actively construct causal shorthands of this kind, adding them onto representations of reality in order to make sense of and predict phenomena whose real nature, in the world, is very complex. That is, in fact, their function: to act as causal placeholders, with associated ontological commitments and inference rules that generate predictions, even if necessarily rough or statistical, of how the system they represent (people or groups of people) will behave, much like the rules of the intuitive mechanics system generates predictions about how billiard balls will behave in a given mechanical context such as a collision. These causal placeholders and associated inference rules work—to the degree that they do—because of the statistical mapping relationship in which they stand to the things they represent, which in turn depends on their design features.

The literature on social cognition and its underlying mechanisms is so enormous that here, as elsewhere, I will not try to review a whole field (see Baron-Cohen et al., 2013; Fiske & Taylor, 2013). Nor will I claim that any particular account that I am offering of the underlying cognitive mechanisms will necessarily end up being the right one—results in this field change quickly. Instead, I will try to paint a picture of the *kinds* of things the mind does in this domain, the kinds of meanings it adds, and what this tells us about mind-world fit. One of the things I will try to show—following on my examples from the last chapter—is that there is likely to be not just a single social ontology but many, each of which adds its own distinctions and interpretations to underlying representations. I will suggest that this challenges the common view of "domains" being partitioned off from each other like separate spaces, such that to the extent a representation is "in" or used by one system, it can't be used by another. Instead, I will try to show that the primary role of representational systems is to *add* features, add distinctions, to underlying representations, framing them in ways that are useful for decision-making and for consumption by other systems—and not to constrain them from being processed by other systems by boxing them off. Moreover, newer systems—more recently evolved ones—can add new kinds of representational features that could not previously be entertained, while leaving all of the inferential abilities of previous systems intact.

This leads to new synergies, new flexibilities, that arise from the interaction between systems. Learning can and does play a huge role in this flexibility, not just because of parameters of inference systems that are adjusted by learning, but also because learning mechanisms can operate on the *outputs* of conceptual parsers, making use of the causally relevant carvings they have made in the raw data coming from the world. While some have claimed that flexibility and specialization are

opposites, I will try to illustrate how the progressive adding of new kinds of representational formats is what gives flexibility to minds, not takes it away.

---

How does information come to be processed as, for example, about "intentions" when intentions aren't present in the input? According to the additivist view I am proposing, this happens via a kind of computational "bucket brigade," with earlier systems, like object parsers, processing the information a little, making relevant distinctions, and adding informational tags based on their inferences—including probabilistic inferences—and then passing this information along to the next system.[1] Somewhere after perceptual input has been carved into objects, animacy detectors use features like motion to categorize certain objects as animate and mark them as such (Scholl & Tremoulet, 2000). Additional mechanisms further subcategorize animates into types, like humans and nonhumans, and mechanisms such as face detectors assign properties, like individual personal identities in the case of humans whose face information has been stored in memory (Kanwisher, 2000). Action parsers carve what an animate object is doing into a sequence of discrete, interpretable actions (Baldwin et al., 2001; Baldwin & Baird, 2001; Zacks, 2004; Zacks & Tversky, 2001). It is in interaction with these processes that intentional inference occurs, leveraging the computational work done by these systems and combining information from them in systematic ways: for example, interpreting the "same" action differently depending on who is doing it, or calling up stored representations of what an individual knows in order to infer the intentions behind what he is doing. I will briefly describe these steps and some of the data that suggest that there really *are* processing steps like these.

Remember that our ultimate goal is to have theories that show how information is processed from the bottom all the way up, without any magical steps between. Though my review will necessarily be sketchy, I hope to give you a flavor of how a sequence of relatively simple processing steps can together lead to fairly complex conceptual processing. I don't mean to imply that each step is *necessarily* simple—animate motion perception, for example, is known to be quite complex—but rather that the brain solves complex problems by decomposing them into more tractable parts, each of whose design features we can study and understand.

[1] It is generally agreed that processing in most domains is hierarchical, occurring in stages, though there are various models of how this occurs (e.g., Davis & Johnsrude, 2003; Felleman & Van Essen, 1991; Holland et al., 1986; Lee & Mumford, 2003; Selfridge & Neisser,1960; Ullman, 2007). We will return to the hierarchical organization of brain processes in more detail in chapter 12.

Let's pick up where we left off. We've already seen, in sketch form, how we arrive at packaged representations of objects: Object parsers scan perceptual data looking for edges and contours and other gestalt features such as shared motion and deliver representations in a format that I called object files. Remember that these representations don't *replace* the original information as if substituting an icon like a cognitive chess piece; they leave the original information (mostly) intact but *tag* it with a kind of frame, additional information that says "all this information belongs to the same object."

After this, the process of animacy detection, or parsing of the world into animates and inanimates, occurs (Leslie, 1994; Scholl & Tremoulet, 2000). There are actually several computational routes to this: via motion, most obviously, but also using static features like shape. It's easy to see why an ability to detect animates could be favored by selection: An organism that cannot tell the difference between a rock and a predator will not do as well in fitness terms as one that can. The reason is because animals do things—they behave—and it is useful to be able to condition one's own behavior on what another object might do. This explains why the selective advantage of animacy detection would not apply to plants, which can't do anything based on the information (except in a very few cases, such as the Venus flytrap, which does have a kind of crude animacy detector based on touch). It also suggests that the animate/inanimate distinction is not made just because it can be—the resulting representations must be tagged as animate or inanimate and then passed on to other systems whose job it is to figure out what to do with them (run away, approach, and so on). And there is a reason to expect evolved cognitive systems to treat the animate/inanimate distinction as a relatively discrete, binary one: You either can move and do stuff, and therefore are potentially worthy of monitoring and treating in a special way, or you can't—though of course there can be a statistical, fuzzy nature to the signal detection problem of animacy detection, with error-management tuning of detection thresholds (Gao et al., 2010; Guthrie, 1993).

Given that the whole point of detecting animates is because they can act—move—it is not surprising that the primary kinds of cues that animacy detection systems use are motion cues. And these systems do not use just any motion cues, but rather those cues that are most characteristic of animate motion, such as, for example, cues that the motion is self-propelled, goal-directed, and contingently responsive to external events (Frankenhuis et al., 2013a; Gergely et al., 1995; Johnson, 2003; Leslie, 1994; Scholl & Tremoulet, 2000).

Detecting self-propelled motion is perhaps the simplest of these tasks. Remember that the intuitive mechanics system generates expectations about how objects will behave in response to mechanical forces like collisions and gravity. Such a system expects mechanical objects to move when they are either being collided with or not

supported in a gravitational field; these ontological commitments are engineered into its prediction rules (Baillargeon & Hanko-Summers, 1990; Leslie & Keeble, 1987; Spelke, 1990). Crucially, self-propelled motion violates these rules. Upon observing an object move by itself when no external force appears to be acting on it, the intuitive mechanics system will either not be able to process what's going on or will generate some kind of error: It will be surprised. This lays the groundwork for the evolution of other detectors that notice when the intuitive mechanics system generates an anomaly. This does not necessarily mean that the intuitive mechanics system evolved prior to animacy detection, but animacy detection certainly could leverage the computations of this other system.[2]

A long line of research, starting with the pioneering work of psychologist Albert Michotte, has shown that motion that violates the ontological commitments of the intuitive mechanics system "pops out" as being something entirely different than the motion of colliding billiard balls (Michotte, 1963). This is true even in babies, as developmental psychologist Alan Leslie has shown (Leslie & Keeble, 1987). You can show a "launching" event over and over again, in which a moving ball collides with a stationary ball, launching it forward. Babies watch this for a while and grow bored: It's entirely consistent with what their object mechanics system says should happen. But when you introduce a spatial gap between the launcher and the launchee, so that they don't make contact before the launchee takes off, the causation looks distinctly different and more consistent with an internal source of motion (Heider & Simmel, 1944; Scholl & Tremoulet, 2000).

Notice that what we're calling "self-propulsion" is actually just the initiation of motion with no apparent cause, where "apparent" refers to those causes that the object mechanics system is able to detect (as we will see, the intentional inference system adds further potential causes, like fear or verbal requests, but these are not yet in the self-propulsion system). Things do, in fact, sometimes move without having been collided with (leaves falling from trees, rustling grass). Relying on apparently uncaused motion can therefore result in lots of false alarms. What is a much more reliable cue to animacy—but also more difficult to detect—is goal-directed motion.[3]

[2] Note that it's not necessarily the case that something *must* be parsed as a whole object before inferences of animacy and agency can be implied. It's quite possible, for example, that we can make inferences about animacy or goals for things that haven't fully satisfied the input conditions of object parsing, or that violate them. In chapters 11 and 12, we'll examine which kinds of architectures might support such robustness (i.e., the ability of a system to produce a computation even with incomplete information).

[3] There is indeed evidence that self-propulsion and agency (capacity to act in a goal-directed way) can be perceptually decoupled in infants. Infants may construe objects as self-propelled without being agents and vice-versa (Csibra, 2008; Luo et al., 2009; Setoh et al., 2013).

You can imagine the added value an animacy detection system could get from distinguishing goal-directedness from the simpler but broader class of apparent self-propelledness (i.e., motion not caused by collision). Peering out over a sea of grass moving in the wind, for example, it would be useful to detect the object that is systematically coming toward you through the grass. Goal-directed motion is, of course, not uncaused at all (nor, for that matter, is self-propelled motion); that is the whole point of paying special attention to it. Later, we will get to how the system *interprets* the causes of goal-directed motion, but first, let's ask how you'd build systems to detect it.

A fairly simple cue, and one that animacy detection systems do indeed use, is contingency, or more precisely, contingent reaction at a distance (since contingency, broadly speaking, is characteristic of all causation) (Gergely & Watson, 1999; Watson, 1985). This actually occurs in perception of non-colliding Michotte displays: Even though there is a gap between the launcher stopping and the launchee starting, the launchee's motion appears to have been caused, at a distance, by the approach of the launcher (it's spatiotemporally contingent on it). In principle, any kind of systematic spatiotemporal correlations like this can contribute to a perception of animacy, and research shows that they do (e.g., if a cockroach stops moving when you stop moving, and then starts again as soon as you move, it's being influenced by you). The most compelling cases of goal-directedness involve specific *kinds* of contingency: You know that you're being stalked, for example, when every time you cross the street or turn a corner, the person following you does the same. And these kinds of goal-directed contingencies produce a strong perception of animacy indeed; they are sufficient to tag an object as animate (Gao et al., 2010; Gergely & Csibra, 2003; Heider & Simmel, 1944; Johnson, 2003; Scholl & Tremoulet, 2000; Watson, 1985). For example, developmental psychologist Susan Johnson showed that babies will categorize even a relatively unalive-looking object like a fuzzy ball as alive if it beeps contingently with the noises the baby is making in a call-and-response sequence. Other cues, like eyes, add to the effect (Johnson, 2003).

The question arises here of where mere categorization ends and further inferences begin: about, for example, *why* the object is doing what it's doing. While processing of stimuli such as these almost certainly proceeds in stages, I want to stress again that even processes that appear "merely" perceptual can, in fact, be conceptually laden. Seemingly conceptual principles of goal-directedness, for example, can and do take part in perceptual inference. The conceptual part need not be explicitly represented anywhere; it manifests in the ontological commitments that are embodied in the inferential procedures that the system uses. Here, it is as if animacy detectors "know" the principles underlying goal-directed behavior, though, as in bacterial phototaxis, it's probably more accurate to say that they merely leverage or exploit these

principles. They are looking for the signal of animacy in the incoming information, and sometimes complex features of the signal have higher predictive value, or what is sometimes called "cue validity," than simpler features (Brunswik, 1955). In building these systems, natural selection has homed in on combinations of properties in motion that are highly diagnostic of animacy, and they are diagnostic because of the causal principles underlying goal-directed behavior.

---

Motion cues are not the only kinds of cues involved in detection of animate objects. There are other mechanisms that use static features—properties that are attached to the objects themselves, like their colors, shapes, and appearance—to categorize objects as animate, and to categorize *within* the category of animates. Sometimes, these operate after motion detection systems, sometimes in tandem with them, and sometimes in the absence of them.

In psychology, there exists a host of theories about how objects are recognized or categorized (Palmer, 1999; Ullman, 2007; Yuille & Kersten, 2006). What most of these theories have in common is that the process of object recognition is broken down into parts. What exactly the subcomponents are is where different theories disagree. We won't worry about that level of detail here. However, all of these theories imply that a substantial amount of information must be assembled in steps. Early on, object parsers have only tagged the object *as* an object; beyond that, representations of different features must be assembled, such as its shape, whether it has parts or structure, and so on. The visual system has to deal with the problem of three-dimensional objects looking different from different perspectives. For categories with multiple members, like dogs or people or furniture, it has to deal with variation within the set. For reasons like these, vision scientists generally agree that object recognition is a hierarchical process in which richer and richer representations are built up, and it takes time—it occurs at relatively "higher" levels in visual processing, or later in time, than some of the more basic processes like object parsing that we have discussed so far. Indeed, there is almost no end to the fine-grainedness with which one can categorize, or make distinctions about, objects based on their properties, and some of these go into very abstract conceptual levels indeed, such as deciding which of two paintings is a genuine Vermeer and which is a forgery. As it turns out, because of the large number of specialized mechanisms involved in visual object recognition, there are a host of bizarre disorders called visual agnosias that can result from brain damage to these mechanisms, impairing things like the ability to decompose objects into parts or to see more than one object at a time, while leaving the rest of vision relatively intact (Farah, 1990).

The complexity of object recognition is one reason why motion detection might be a quicker shortcut to parsing between animate and inanimates than taking the time to figure out *what* the thing is, which usually occurs a bit later (we see something move in the grass, jump back, and only then realize it's a snake). It might also be why, as I mentioned in the last chapter, infants appear to develop the ability to track objects spatiotemporally before they begin to rely on the static features of objects, because developing a representation of an object's features (shape, color, texture, parts, category) not only takes time but requires many mechanisms to have developed.

That said, however, our visual systems can and do detect animates based on static cues alone, and several lines of evidence suggest that there are mechanisms that have evolved specifically *for* detecting animals quickly on the basis of static cues. One set of studies, by psychologist Josh New and colleagues, makes use of a phenomenon called "change blindness" (New et al., 2007a; Simons & Levin, 1997).[4] As you might expect, what we pay attention to is very important to what we can and do notice—but it actually matters even more than you might think, because you're paying attention to a lot less than you know. Although we have a very wide visual field—about 60 degrees of angle on either side of the eye—and are thus in some sense "aware of" or represent information across that wide field, we are only actually focusing on or paying attention to a very tiny bit—only a few degrees right in the middle, an area known as the fovea—at any given time. And it turns out that events that occur outside this area are relatively hard to notice, even though light bouncing off of them does land on our retinas and is actually processed. This probably reflects a quantity-quality tradeoff of sorts: a relatively small but high-resolution processing area for what you know you need to attend to now (or what your brain estimates is the current fitness-best thing to look at), surrounded by a relatively large but low-resolution area to cast a wide but crude net for early detection of "incomings."

As an organism proceeds through the world, things are happening all around it, and mechanisms in its brain have to decide what to attend to, where to focus the spotlight of attention. What psychologists have shown in a host of clever demonstrations is that things are happening around us all the time, change is occurring, and we don't notice much, if not most, of it. In one study by Daniel Simons and Christopher Chabris, subjects watched a video of a ring of people passing a ball around and were asked to count how many times the ball was passed. While they were following the

[4] Change blindness, or failure to notice a change in the perceptual environment, is related to "inattentional blindness," which is defined more broadly as failure to notice a stimulus that is present in the environment (Mack & Rock, 1998). The gorilla suit experiment described below is often considered an example of inattentional blindness, though it also arguably involves failure to notice a change in the scene.

ball, a person in a gorilla suit entered the circle, made a chest-pounding display, and left. In most conditions, fewer than half of the subjects reported seeing the gorilla, and sometimes as few as 8% (Simons & Chabris, 1999). In another study, a researcher approached a person on campus and started talking. The conversation was interrupted by people carrying a large object between the researcher and the subject, and after they passed, the researcher had been replaced by a different person. About half of the subjects tested didn't report noticing they were now talking to a different person (Simons & Levin, 1997).

The usefulness of this kind of research—aside from perhaps disabusing some of us of cherished ideas about our own perceptiveness, or making us think twice about texting while driving—is that investigations of what kinds of changes we *do* notice, and which ones we don't, can reveal interesting design features of cognitive systems: things they can and can't detect, and their priorities for allocating attention. New and colleagues used such a design to test what they call the "animate monitoring hypothesis": that there is a mechanism that preferentially allocates attention to animate things in a scene, because these are most likely to move and thereby warrant some action on our part. In a series of studies, they altered the locations of objects in natural scenes and measured when people noticed the changes. They found that people detected changes to the people and animals (i.e., animates) at a much higher rate, and with quicker reaction times, than changes to any of the inanimate objects—even when the inanimate objects were large, such as an entire building changing location (New et al., 2007a). This implies the existence of detectors—in this case, for animates—that are constantly monitoring the world for us and telling us what to attend to. And in this case, the detectors must use static cues to determine what is animate and draw our attention to it because the pictures were static.

Other studies provide evidence for such detectors as well, exploiting the distinction between serial and parallel processes. Processes requiring attention are typically serial, in that attention must be focused on one item or task at a time, whereas processes not requiring attention can occur in parallel, and thus sometimes more rapidly (Treisman & Gelade, 1980). One study asked subjects to either count how many snakes there were in an array of flowers or count the flowers in an array of snakes. If presented with five snakes randomly distributed among 25 flowers, you see all five snakes in an instant, a phenomenon known as a "pop-out" effect—a good sign that a specialized detector is at work. If you look for five flowers among 25 snakes, though, it is the snakes that still pop out; you have to ignore them and count the flowers individually, which takes much longer (Öhman et al., 2001). Other studies have looked for detectors in far peripheral vision—the low resolution zone where it's hard to tell the difference even between a triangle and a square (try it: Have someone move objects slowly into your far peripheral vision, without looking directly at the

objects, and see how long it takes you to identify them). In this region, people are able to detect the presence of animals like tigers and elephants far better than inanimate objects: Again, this is consistent with the idea of animate detectors and with the idea that peripheral vision acts as an early warning system for things that are fitness-important (Braun, 2003; Thorpe et al., 2001).

Together, these studies suggest that there may be something like perceptual "templates" for animates. The idea of a template is something that makes many psychologists nervous, perhaps because it implies the existence of some kind of innate picture of living things to which inputs are compared. However, as I hope I've stressed enough by now, informational structures do not need to be picture-like to do their job. Object parsers, for example, do not compare inputs with a picture of a generic object to determine whether it's an object or not—what would such a picture be? Instead, they have algorithms that reliably detect features that are characteristic of objects. While we don't yet know precisely what features detectors for animates are using, the idea that detectors are operating is hard to escape.

As I'll discuss in greater detail later, natural selection can create development processes (which I'll call module-spawning processes) that can give rise to specialized structures like these detectors without having to specify every detail of what the finished template will actually look like. Many of the requisite details can be filled in by experience using specialized learning mechanisms. And indeed we might expect them to be designed that way, because the world is a useful source of information about itself. In other words, a template-spawning system might instantiate the inductive bet *that* the world contains animates, and that they are worth detecting quickly and peripherally. It might also contain some initial procedures for detecting them using particular cues (e.g., motion), as well as procedures for building more detailed templates using those cues, which could be thought of as the Bayesian priors of the system (see, e.g., Moya, 2013).

This will be a kind of general model of development that I will propose frequently operates in evolved systems: Natural selection can provide the developmental materials to produce specific *types* of outcomes, like animal templates, without necessarily specifying the details of each *token*. Instead, natural selection provides the spawning system that allows the tokens to be built on the basis of experience. For example, we might have an artifact system that allows us to acquire a concept of forks. The "type" would be *artifact*, or perhaps more specifically handheld *tool*. The learning system would make specific inductive bets about tools, such that they have a function, can be manipulated, have a characteristic shape, etc. It might also be prepared to be wired to the visual system in a way that allows us to see and learn about tools and wired to the motor system in order to allow us to grasp and manipulate them (Culham & Valyear, 2006; Johnson-Frey, 2004). However, the specific token, fork, would not

be in any way "specified" by the system; the details of forks would be learned. Such a system would be domain-specific—specialized for tool learning—and yet flexible, designed to interact with other systems, and only "innately specified" in terms of its learning and organizational procedures. Such learning designs gain power through their inductive bets—they can solve frame problems and needn't discover certain principles, such as that tools can be manipulated—and yet they are also quite flexible *within* the domain in which they are designed to operate. Systems like these could be the norm, rather than the exception, in brain evolution.

---

As I alluded to, the kinds of distinctions and discriminations we can make with respect to objects are almost limitless, ending in distinctions as subtle as distinguishing a frown from a smile, a wink from a blink. We will get to how we parse and interpret actions, but first, let us consider how we carve entities into categories *within* the domain of the animate.

I already mentioned the possibility of templates for certain kinds of organisms, like snakes, spiders, and humans, with the possibility of spawning new template tokens and adjusting their properties via experience. These categorizers perform an important service in the cognitive division of labor: They do the job of deciding what the thing is and then tag it with that information, so that future systems dealing with the object no longer have to worry about finding out what it is. Once one system has done the job of checking and says "yep, that's a lion," they pass the results of their decision along, bundled into the representation, for other systems to use. This adding of informational tags as representations work their way through the system is a key source of flexibility due to cognitive division of labor, just as economies become more flexible when not every worker is doing exactly the same thing (Barrett, 2005b).

One critically important kind of animate to identify is humans. It turns out that we can detect humans through a wide variety of means, such as voice, shape, and motion, even using minimal cues like points of light attached to the limbs of someone walking in the dark (Belin et al., 2000; Blake & Shiffrar, 2007; Chang & Troje, 2009; Johansson, 1973; Kanwisher, 2010). In addition to regions that appear to be specialized for perception of whole body motion (Grossman et al., 2010; Thompson et al., 2005), there are certain parts of the body for which our brains contain specialized detectors, including faces (Kanwisher & Yovel, 2006) and hands (Iacoboni et al., 2005). This makes sense given the importance of these body parts as sources of information about a person's emotions, intentions, and actions.

There is evidence for multiple specialized and evolved mechanisms for extracting information from the face, each solving a particular adaptive problem. There is the

problem of detecting that something *is* a face. There is the problem of face recognition or inferring the identity of the person (in cases where the person is known) from facial features. And there is the problem of inferring emotion and other mental states from facial expressions. There is evidence for specialized mechanisms that solve each of these problems. These mechanisms are worth briefly visiting because of their importance in human social interaction and also because of what they might tell us about the design of evolved systems that are specialized, yet specialized to deal with inherently variable content: every face you could possibly recognize.

Face recognition has been a topic of both extensive study and extensive controversy in psychology. It is a good example of a case where the very same ability is thought by some to be the result of a specialized, evolved system, and by others to be the byproduct of a more general-purpose system. The more domain-specific claim is that there exists a system evolved specifically *for* face recognition, which functions to identify individual faces and to tag them with the identity of the individual, thereby allowing all of our stored information about that individual to be associated with their face (Kanwisher & Yovel, 2006). The more domain-general claim is that face recognition is just a particular instance of expertise with objects that we see a lot—faces—without there having been any selection specifically *for* face recognition (Gauthier & Nelson, 2001). We'll return to this debate in a moment.

If a specialized face recognition system exists, what does it have to do, informationally speaking? First, it has to deal specifically with faces. As I mentioned above, this might be done in a two-step process where one process infers that an object *is* a face and a second process infers *whose* face it is. Like snake detection or spider detection, face detection could proceed via a template that abstracts general properties that are typical of all faces. Although there remains controversy about this proposal as well, there is evidence that a minimal template of this kind develops very early in infancy (Morton & Johnson, 1991). What are the features that the minimal template uses to discriminate faces from non-faces? Infants' face detection template was originally thought to be something like a schematic pattern of two eye blobs and a mouth blob, based on findings that such a pattern drew newborn babies' attention but a scrambled pattern of the same shapes did not. Recent findings suggest that the earliest template could be even more minimal, perhaps just a preference for configurations with more elements in the upper than in the lower half (Macchi et al., 2004).

Once faces are detected, discriminating among the faces of individuals is a much more difficult task. While the face detection system can tag an object as a face and thereby admit it for processing by a face recognition system, the presence of features that are used to detect faces can't be used to discriminate among them, since all faces will have those features. The face detector could, of course, tag certain relevant

features that it used in making its decision—*here* are the eyes, *here* is the mouth—and pass these along to the facial identity detector for help in making its decision. However, it is clear that the system must be designed to learn to recognize individuals; nobody would claim that there is an innate template for every face you come to recognize.

It is generally agreed that the combination of features used to discriminate between individual faces must be complex. It is sometimes called "configural," "holistic," or "global" processing, because what varies among faces is not just the presence or absence of certain features, but how they are arranged (Maurer et al., 2002).[5] We still do not know how many of the relevant dimensions of variation are provided as ontological commitments of the learning system, and how many it has to extract from its experience of the variation it sees, but it is clear that the system has to somehow carve the spectrum of faces up along relevant dimensions, process them configurally (in terms of spatial relationships within the whole object), be able to do this from multiple perspectives (e.g., front, side), and create an individual identity file for each individual the system can recognize.

There is evidence that each of these adaptive problems is solved by the face recognition system. Some of this evidence comes from typical signatures of face processing that you can easily demonstrate to yourself at home. Perhaps the most well known of these is the "inversion" effect, where faces are much more difficult to recognize upside down than right side up (Diamond & Carey, 1986). This is good evidence for configural processing, because objects that just look for individual features, as opposed to their arrangement (e.g., blue eyes, black hair) have much less difficulty with inversion. You can still recognize a zebra when it's upside down, for example, from the stripes and overall shape, but telling the difference between two faces upside down is much harder (this is also evidence that the facial template in the visual system has an orientation, right side up, because of the way we typically view faces). Consistent with the configural processing idea, manipulating individual features within the face when it is upside down—such as individually turning the eyes or mouth right side up—is hard to notice, as if we are not processing those things as individual objects. Turn the face right side up again, though, and the now-upside-down eyes and mouth become hideous. This is sometimes called the "Margaret Thatcher illusion" because the first experiments used her face (Bartlett & Searcy, 1993).

[5] Note that configural processing also appears to occur in some other cases, such as processing the complex motions of animal bodies (Chang & Troje, 2009; Thompson et al., 2005). It seems likely that the specific details of configural processing differ across these domains, though this remains an empirical question.

Much of the evidence for the design features of face processing, though, comes from what happens when it breaks down. Prosopagnosia is the technical term for the inability to recognize faces, and it can result either from brain damage to the areas involved in face processing or the failure of those areas to develop properly in childhood (Farah, 1990). An interesting aspect of prosopagnosia, and many other forms of brain damage as well, is that it damages one cognitive ability while leaving others intact. This is known in neuropsychology as a dissociation (Shallice, 1988). For example, many prosopagnosics can tell that something *is* a face; they just can't tell who it is, suggesting intact face detectors but broken identity detectors (de Gelder & Rouw, 2000). There have been many detailed studies of people with prosopagnosia that look at where the boundaries of their deficits lie. If, for example, face recognition is actually handled by a more general-purpose system, then when that system is damaged, the effects should appear for more than just faces. Recently, psychologist Brad Duchaine and colleagues conducted a study in which they cataloged various "byproduct" theories of face recognition—theories that propose that it is a result of a more general ability—and tested one prosopagnosic man, Edward, using stimuli that should, on these other theories, result in processing difficulties. Edward was normal on every test except for faces, suggesting that it's possible to have impairment of face recognition while leaving more general- purpose abilities, such as configural processing more generally, intact (Duchaine et al., 2006).

Critics of the domain-specific view of face processing have suggested that certain other kinds of stimuli—notably, complex stimuli that seem to require some kind of configural processing—can exhibit the same cognitive signatures as face processing, such as the inversion effect. One study, for example, found the inversion effect for dog experts trying to discriminate between dog breeds while upside down (Diamond & Carey, 1986). Other studies using artificial stimuli called "greebles"—biological-looking but not exactly face-like––were designed to show configural processing. Psychologist Isabel Gauthier trained subjects to become "greeble experts" by rewarding them for distinguishing between individual greebles. She found not only signatures like the inversion effect, but also activation of face areas in the brain. She concluded on the basis of this that these areas are not, in fact, specialized for faces per se, but some broader class of objects (Gauthier & Tarr, 1997).[6]

However this debate ends up playing out empirically, it illustrates the potential importance of distinguishing between the proper and actual domains of cognitive mechanisms, where the proper domain refers to things the mechanism was selected

[6] Several studies show that faces and non-face objects are processed differently in the brains of infants, suggesting that faces are not merely treated as a subset of objects, even early in development (de Haan & Nelson, 1999; Southgate et al., 2008).

to process, and the actual domain refers to things it *can* process by virtue of its features (Sperber, 1994). Importantly, the actual domain will *always* include entities that were not in the proper domain, because any set of evolved detection criteria can be satisfied by stimuli that were not present in ancestral environments but that nevertheless satisfy its felicity conditions, as demonstrated nicely by nest parasites and Lorenz's geese. For any evolved system, it's virtually certain that stimuli can be found that will be processed by the system even though they are not part of its proper domain. The plethora of visual illusions that trick our visual systems are an example. They are "unnatural" stimuli that are nevertheless processed by visual mechanisms using their normal inductive bets, leading to inferences that are in fact wrong. In other cases—possibly in *most* cases—evolutionarily novel stimuli might be processed by evolved systems in perfectly normal and adaptive ways, because it is not a coincidence that they share features with ancestral stimuli.

The proper/actual distinction is important to the face recognition debate because the existing theories differ on what they hold the proper domain of the underlying system(s) to be. If it's faces, then it's possible that certain kinds of face-like objects—such as greebles, or the fronts of cars, or computer-animated faces of imaginary animals or robots—could be processed if they share enough features with ancestral faces to be part of the actual domain. Crucially, if this is true, then evidence for byproduct processing of this kind wouldn't necessarily falsify the face-specific hypothesis—bad news for testing between the hypotheses. If, on the other hand, the proper domain of the underlying mechanism is objects more generally, then we'd want to find out what the system's underlying inductive bets are, and whether they truly span the entire realm of objects, as opposed to being more narrowly targeted. As we'll see later, there is evidence that module-spawning systems can indeed develop ontogenetic specializations for evolutionarily novel categories of objects, such as written characters—but this is probably not a coincidence, as writing systems have evolved to satisfy the inductive bets of object recognition systems (Changizi & Shimojo, 2005; Dehaene, 2009).

No matter how it turns out, the debate over face recognition is a useful model for testing between alternative hypotheses in evolutionary psychology (Ketelaar & Ellis, 2000). Neither the face-specific nor object expertise hypothesis is non-evolutionary, in principle. Nor do either fail to specify a domain: faces and objects, respectively. They are merely different proposals about the nature of the underlying specializations, and once these proposals are properly specified in terms of expected design features and their empirical signatures in psychological data, we can use the ordinary methods of psychology and neuroscience to test between them.

We are to the point in information processing where, upon seeing a person, the brain has identified the person as an object, identified the object as a person, and identified the specific person that it is. Unrecognized people are still recognized *as* people, of course, and we also assess them in terms of many other characteristics aside from their individual identities. For example, we can categorize their gender, age, race, and assign them to other social categories like profession or class based on how they are dressed. Social categorization is complex and involves many kinds of cues and inferences (Fiske & Taylor, 2013).

In addition to categorizing people, we attempt to make sense of what they do: We interpret their behavior for the purposes of our own decision-making. Just because I've presented them in this order, I don't mean to imply that you categorize people first and then interpret their behavior—though the processing can sometimes happen in this order. Instead, it's better to think of these as processes that *interact*, such that knowing who a person is helps with the interpretation of their actions and vice-versa. These interactions are sometimes known as context effects, and though they are widely held to be uncharacteristic of so-called modular systems made of specialized parts, we'll see that this view isn't necessarily correct.

A critical step in interpreting the behavior of others is sometimes called action processing or action parsing (Baldwin et al., 2001; Baldwin & Baird, 2001; Blake & Shiffrar, 2007; Decety & Grèzes, 1999; Gallese et al., 1996; Newtson, 1973; Sommerville & Woodward, 2005; Woodward & Sommerville, 2000; Zacks, 2004; Zacks & Tversky, 2001). Parsing, you will recall, refers to carving a continuous stream of information into meaningful units. As elsewhere, what might appear to be a single adaptive problem actually consists of many separate ones. There is the problem of deciding whether some aspect of an agent's movement is an "action" at all: Did that person touch you on purpose, or did their hand brush you by accident? Then there is the diversity of the different "acts" a person can perform, ranging from running, to picking something up, to dancing, to shrugging, to saying something, to smiling. It is hard to imagine that there is a common set of cues that all of these acts have in common. This is a good reason to suspect that the system operates via division of labor, carving movement into chunks and refining them bit by bit in a processing chain, with more conceptual carvings of actions (e.g., categorizing them according to goal) emerging only at the end of the chain.

The study of action parsing is a rapidly growing field (Johnson & Shiffrar, 2013; Rutherford & Kuhlmeier, 2013). This growth is spurred in part by the realization that there are deep frame problems that need to be solved every time we look at someone else and try to figure out what they're doing. And even when one is aware that frame problems exist, understanding how they are solved—enough to, for example, create a computer algorithm to parse actions from video—is difficult (Moeslund et al., 2006).

But there have been a variety of proposals of how it might work. Here as elsewhere, we know some things about the kinds of adaptive problems the system has to solve, and does solve, so we can use this to narrow down the possibilities for how it might do it.

Psychologist Dare Baldwin and colleagues conjectured that if the function of action parsing systems is to carve the action stream into chunks that are relevant for interpreting agents' goals, then the chunks they produce should fall at the boundaries of where goals are completed. To test this, they used a method originally developed to find the edges of units that speech parsers produce in our perception of speech. As it turns out, boundaries between words don't exist in the sound stream of spoken speech, even though we perceive the words as distinct. Instead, parsers are chopping the speech up into the units of meaning that we perceive, painting boundaries onto continua. Consistent with this, listeners are much more accurate at judging when a beep occurs if experimenters place the beep at a boundary between two perceived words rather than in the middle of a word (Fodor & Bever, 1965).

In a modification of this method, Baldwin and colleagues showed subjects videos of actions, such as a woman opening a refrigerator door, and played beeps at various points in the video: for example, in the process of reaching for the door handle, or right at the moment it was grasped. When they replayed the video later without sound and asked subjects to press a button at the points where beeps had occurred, they found that subjects were much more accurate in remembering the location of beeps when they occurred right when an action was completed (Baird & Baldwin, 2001). Similar results were found in 10–11-month-old infants, using a slightly different method, suggesting that what might seem to be a continuum of movement is actually being carved into units of meaningful action, where "meaning" is related to completed goals like grasping, opening, lifting, and drinking (Baldwin et al., 2001). Parsing of this kind occurs in many domains of human action, from whole body movements, to limb movements, to facial expressions (Adolphs, 2002; Jusczyk & Luce, 2002).

If action parsing is occurring, there is likely to be a logic to it: some set of underlying principles or ontological commitments that the system uses to make its carving bets. There are many possibilities for what such action logics might be, and how specific they are to particular kinds of action. It's possible, for example, that the logic is no more specific than a statistical analysis of whatever inputs the brain receives; or, there could be distinct parsers for speech, actions, objects, etc. (Kirkham et al., 2002; Marcus et al., 2007; Saffran et al., 1996). If domain-specific parsers exist, then they should exhibit at least some inductive bets that are specific to their respective domains. With respect to action, developmental psychologists György Gergely and Gergely Csibra (2003) have suggested that infants possess an action interpretation system that embodies principles of "rational action" that agents might be expected to

(usually) obey. One such principle might be efficiency: minimizing energy expenditure by taking the shortest path or means to a goal.

In one of several tests of this idea, they showed infants a video display of a dot that displayed animacy cues—self-initiated expansion and contraction, like breathing—and then moved toward a barrier and "jumped" over it, thereby reaching another dot. I put "jumped" in quotes because that is the semantically colorized version of what actually appeared on the screen; the motion trajectory gives a strong impression of one dot "trying" to reach the other dot. Did infants parse the event this way as well? Gergely et al. found that when the barrier was removed, infants expected the dot to take a straight-line path to the other dot, instead of jumping. In other words, they were less surprised if the dot moved in a straight line when there was no wall than they were if the dot continued to jump for no "reason" (again, the colorized interpretation). This suggests that the infants had encoded the dot as an agent with a goal, and assumed that agents typically take the shortest path to their goal, not making unnecessary leaps in the absence of barriers (Gergely et al., 1995).

Although the idea that parsing mechanisms solve frame problems by embodying ontological commitments about the logic of action makes sense, it's possible that what appears to be a global "rationality" assumption is really composed of what might be called "islands of competence": smaller schemas of action logic, each tailored to specific chunks of action space that are characterized by shared principles (Frankenhuis & Barrett, 2013). Whole organism motion along spatiotemporal paths might be one of these, where principles like "shortest path" can be instantiated in the rules of an inference system much like gravitational principles can be engineered into intuitive mechanics systems. Another might be the action logic of specific body parts and what they do, including, for example, hands and faces. Evolutionarily, this is not an implausible proposal, because of the recurring importance of hands in human action and the fact that grasping is one of the main things that human hands are designed to do. If you wanted to build a baby who could figure out what other people were doing, building in some implicit assumptions about hands might not be a bad idea.

There are several lines of evidence for hand-specific action parsers that embody inference rules about what hands tend to do. One thing hands do is to reach across space toward desired objects, with fingers open and palms facing the object, and then grasp the object in order to manipulate it. The grasping action of hands is indicative of goals, and the goal of a grasping hand is typically related to what is being grasped. If I see you reaching for the chocolate cupcake and not the vanilla one, I can make an inference about your preferences. Developmental psychologist Amanda Woodward and colleagues have conducted a series of studies in which infants observe hands grasping objects, and use a violation-of-expectation paradigm to assess what infants conclude about the meanings of these events. In one study, children were shown a

puppet-show stage with two objects, a ball and a bear. A real human hand emerged from one side of the stage and grasped one of the two objects (e.g., the bear). After being shown this, children were surprised if the hand later reached for a different object (e.g., the ball), even if it was in the same *location* on the stage where the hand first reached. They were not, however, surprised if the hand reached again for the bear, even if it was now in a new location and it had to reach over or around the ball to grasp the bear (Woodward, 1998). This suggests that the infants encoded the hand as having a preference and assumed it wasn't just reaching for, for example, the closest object.

What is particularly interesting is that this effect appears to be specific to grasping hands. When a mechanical claw is shown reaching for an object, babies do not encode it as wanting a specific thing. This is evidenced by the fact that they are not surprised if it reaches for different things later. Moreover, when babies see a hand reach out and the *back* of the hand touches the bear, rather than grasping it, the babies do not encode brushing against the bear as the specific goal of the hand; they are not surprised if the back of the hand later touches the ball (Woodward, 1999). This suggests that even though the babies are undoubtedly doing a spatiotemporal analysis of the scene, it's not *just* the "low-level" statistical properties of it, such as what is touching what and for how long, that are influencing infants' responses. Otherwise, they would respond the same way in the claw, back-of-hand, and grasping conditions (see also Leslie, 1984).[7]

Finally, much like Gergely et al.'s babies who encoded jumping as a means to an end, babies encode steps on the way to grasping as means to ends rather than goals themselves. Woodward showed that a hand opening a lid on a clear plastic container to grasp what's inside is not encoded as wanting the lid or having a preference for the container itself, but as wanting what's inside. If an infant sees a hand open a clear pink-tinted container to reach for a bear inside, the infant is not surprised when the hand later reaches for a green-tinted box that contains the bear, ignoring the pink-tinted box that now contains something else (Woodward & Sommerville, 2000).

[7] While it seems plausible that grasping is treated as a special kind of event in action parsing—and there is brain evidence for specialized perception of grasping (e.g., Iacoboni et al., 2005)—this does not mean that infants would be unable to interpret goals involving, for example, touching objects with the back of the hand, if context is provided that makes sense of this touching. For example, Király et al. (2003) show that infants can interpret actions using the back of the hand, such as pushing an object, in a goal-directed fashion. Moreover, preferences might be attributable through a variety of agentic cues, not just grasping. Luo and Baillargeon (2005) have shown that five-month-old infants can attribute preferences to a moving box in a Woodward-like scenario where the self-propelled box shows a consistent tendency to move toward one of two objects.

These results extend to imitation. Developmental psychologist Andrew Meltzoff showed that infants will imitate actions being performed by human hands, but not the same actions being performed by a mechanical claw (Meltzoff, 1995; Slaughter & Corbett, 2007). Moreover, a variety of studies have shown that for incomplete actions (e.g., a hand reaching to grasp a ball but not being able to reach it), infants do not imitate precisely what they observe but instead produce the completed action, even if it has not been observed (Carpenter, Call, & Tomasello, 2005; Hamlin et al., 2008; Meltzoff, 1995).

Studies of imitation in young children are increasingly demonstrating its inferential nature and the importance of action parsing and action logic in how children decide to imitate. An early study by Meltzoff used a light box, a box with a large translucent plastic button on top, to show that when an infant sees an adult leaning over to depress the button and turn on the light with her forehead, the child will imitate this action with his own forehead and remember it as the appropriate action to perform on the box even after a week's delay (Meltzoff, 1988). György Gergely and colleagues showed that when babies observe the forehead demonstration, they only perform an exact imitation of what they observed—using their own foreheads to turn on the light—when the adult demonstrator's hands were free when she pressed the button with her forehead, implying that there had been some reason *not* to use their hands. In contrast, when an adult uses her forehead to turn on the light and has her hands occupied, children tend to use their hands to turn on the light instead (Gergely et al., 2002). Not only does this show imitation of an inferred goal rather than exact copying, it shows that babies assume that hands are the primary manipulators of objects and that if other body parts are being used, there must be a reason.

Research on what are known as mirror neurons has begun to examine the possible neural substrates of action parsing and imitation (Gallese et al., 1996; Iacoboni & Dapretto, 2006; Rizzolatti & Craighero, 2004). Mirror neurons are a type of neuron in the motor area of the brain that fires both when the individual performs an action and when that individual observes another person performing the same action, or rather, the same type of action, such as reaching and grasping. The discovery of these neurons has led to a great deal of excitement, because if they help to parse the actions of self and other into common underlying types, they could play an important role in imitation and social learning. Interestingly, mirror neuron researchers have shown that the mirror neuron system is parsing the action stream in terms of goals, and not just based on raw perceptual similarity. Neuroscientist Marco Iacoboni and colleagues put subjects in a brain scanner, where he could record their mirror neuron activity, and had them watch films of a hand reaching for the handle of a coffee cup and grasping it. Crucially, in one scene, the hand was reaching for a cup on a table full of dirty dishes, as if cleaning up the table after a meal. In another scene, the hand was

reaching for the cup on a table set for a meal, as if to take a first drink. In both cases, the physical motion of the hand—reaching for a cup—was identical, but the implied goal was different (cleaning up versus drinking). The results showed that subjects' mirror neurons responded differently for the two different goal types, suggesting that the mirror neurons can't just be responding to the mere perceptual properties of the reaching motion, but rather to its goal (Iacoboni et al., 2005).[8]

Results such as these support the idea of a parsing system that embodies principles specific to the actions of hands and the way they are used instrumentally to achieve goals. I've used hands as an example, but there must be parsing systems for all of the various actions and gestures we're able to understand, and systems for assembling sub-actions into larger action schemes. For example, there is evidence for brain areas specialized for parsing mouth movements, areas specialized for parsing facial expressions of emotion, and areas specialized for processing whole body movement (Adolphs, 2002; Grossman et al., 2010; Rizzolatti & Craighero, 2004). Action parsing must be present in all species with visual systems that are able to categorize others' actions in some way, from pigeons to people, but the *kinds* of actions species can distinguish are presumably quite variable. The distinctions a species can make are likely to tell us a lot about their social life, and the purposes for which they attend to what others are doing. In the case of humans, the number of action types we can distinguish is enormous, and we clearly have the ability to learn to parse actions that are complex and evolutionarily novel, from ballet moves to the gestures of sign language—something that we'll have to consider in light of types and tokens and the nature of developmental systems that build our parsers, to which we'll return in chapter 6.

In case you're tempted to write off action parsing as just a few nitty-gritty details of perception, consider this: All of our understanding of human action, from our ability to understand the communicative acts of others to our ability to acquire skills and habits by observing others, must pass through this stage of processing. Action parsing is what paints the stream of action with the meanings we perceive. Evolutionary psychologists Leda Cosmides and John Tooby coined the term "instinct blindness" to refer to a phenomenon that's been recognized at least as far back as

[8] Note that there is a debate about the nature and function of mirror neurons (Csibra, 2007; Hickok, 2009). The name "mirror neurons" is evocative and implies a primary function having to do with comparing one's actions to others' actions, and therefore possibly imitation. However, the empirical data only demonstrate the circumstances under which these neurons fire, and therefore do not rule out a number of other possibilities for their function. By analogy, there must be neurons that fire whenever a particular concept is instantiated—for example, when you say "cow," when I say "cow," when I think about a cow, or when I see a cow. Such neurons could play a role in imitation but it needn't be their primary function.

Plato's allegory of the cave: Because our minds paint the world with meanings, and because these meanings are the very substrates of thought, we can easily be unaware that the meanings aren't coming from the stuff itself, but from us, or rather, from the evolved mechanisms in our brains (Cosmides & Tooby, 1994). These are the mechanisms that turn the muscular activity of a mammalian face into a smile, the locking of someone's arms around your body into a hug, the pressing of mouths against each other into a kiss. Without them, social life would be meaningless. To borrow from Shakespeare, it would be a series of sounds and motions, perceivable perhaps, but signifying nothing.

My intention here has not been to give an exhaustive sketch of social ontology. We will continue to explore it throughout the book, and even then, my account will be necessarily incomplete. Instead, what I've tried to do in this chapter is to expose the tip of the iceberg of what we'll need to explain about human social cognition. Even this very small sliver of cognition, I hope, demonstrates a few basic points about what that explanation will need to look like. First, I hope you'll agree that it's becoming increasingly implausible that all of human cognition will ultimately be accounted for by a few general-purpose mechanisms. We haven't even gotten past the stage of decoding a handshake and already we're dealing with the interaction of many specialized systems. Second, I hope you're beginning to see that in order to account for even simple, lower-level systems such as this, the standard nature/nurture, innate/learned dichotomy is probably not going to cut it. Clearly, a system that takes the incoming stream of perception, slices and dices it, and packages it into units of meaning involves some elaborate organization, and that organization can't be entirely learned anew by every individual's brain via trial and error. Something has to be built into the design of the developmental machinery in order to produce this organization via interaction with the world. On the other hand, the idea that everything about this machinery is innate is equally untenable. There can't be innate parsing templates for kisses and hugs, nor can your genome contain the innate specifications for all the faces you'll be able to recognize, including those of people who haven't even been born yet.

Instead, the kind of theory we're going to need is going to have to account for both sides of the specialization coin. First, it's going to have to account for what's sometimes called the "reliable development" of evolved systems across individuals: mechanisms like object parsers, action parsers, agency detectors, face recognizers, and the higher-level organization of these mechanisms into orchestrated systems of parts that interact by design in a coordinated fashion (Barrett, 2006; Tooby & Cosmides, 1992). This cannot occur entirely by genetic specification, but rather via interaction between developmental systems, including the genome, and the developmental environments in which those systems build the phenotype.

Second, our theory will need to account for the degree and kinds of variability that we see in each of these systems. It will have to account for the fact that the face recognition system is able to learn to distinguish between thousands of faces, that the action parsing system can learn to interpret an untold number of actions, and so on—and yet, that each system does so with its own logic and ontological commitments.

This two-sided coin of specialization and flexibility applies to most if not all of the brain's mechanisms. There are not the innate mechanisms and the learned ones, the lower-level mechanisms that are the result of evolution by natural selection and therefore specialized, and the higher-level ones that are the result of learning and therefore flexible. Instead, we'll need to think of cognition as composed entirely of specializations, all the way up—but specializations that are diverse, and designed to deal with a world that is constant in some ways and variable in others.

# 5

# MINDS

One of the most remarkable things about humans is the degree to which we get inside each other's heads. I don't mean that literally, of course. What I'm referring to is how much the contents of one person's mind can influence the contents of another's. For example, what used to be the contents of my mind—these words—are now, at least momentarily, the contents of yours.

There are many ways in which this mind-mind influence occurs and many mechanisms involved, some of which we've seen already. For example, we've seen the importance of goals—or more precisely, inferences about goals—in the understanding of others' actions, and in our ability to learn from them. Much of this could be unconscious: We might not be aware that when we see a turn signal blinking on a car, we're making an inference about another's goal. But we are making such an inference, or at least, something in our minds is. And many of the ways in which one person's mental contents influence those of another are even more indirect. For example, investors' beliefs about the state of the economy influence stock prices, which in turn influence others' beliefs, goals, and behaviors. One person's choice of clothing influences another person's, transferring a preference from mind to mind.

When we are actually making inferences about what other people think, whether we are consciously aware of it or not, specialized mental machinery is involved. The ability to make such inferences is sometimes called mindreading or theory of mind. There are in fact many terms for this ability, and nearly as many theories of how it happens (Baron-Cohen, 1995; Heider, 1958; Jones et al., 1972; Nichols & Stich, 2003; Tomasello et al., 2005; Trevarthen & Aitken, 2001; Wellman, 1990). Some theories involve more complex mechanisms and inferences, and some less. Intersubjectivity, for example, is a term that is sometimes used to refer to the entanglement and mutual influence of two or more individual's subjective (mental) states, without necessarily implying inference (Trevarthen & Aitken, 2001). Some scholars claim that certain forms of mindreading, such as empathy, do not require inference, because some mental states can be directly perceived in actions and expressions (Gallagher, 2007; Zahavi, 2011). Among those who subscribe to the idea of inferential mindreading, many are concerned with the difference between

what is sometimes called "behavior reading"—sensitivity to externally observable cues like facial expressions without attribution or representation of mental states per se—and "true" mindreading, which involves internally representing others' mental states, such as their beliefs, desires, and goals (Heyes, 1998; Penn & Povinelli, 2007; Povinelli & Vonk, 2003; Whiten, 1996). Even more narrowly, some people reserve "theory of mind" to refer specifically to the attribution of beliefs, not the attribution of goals, dispositions, or emotional states (see, e.g., Bloom & German, 2000, on the widespread assumption that theory of mind equals the ability to attribute false beliefs in particular).

Here I am going to take the broadest possible view of what mindreading is and what it might be. In accordance with the Shannon view of information I laid out above, I'm going to call anything mindreading that uses a cue or cues that index another's mental state: for example, eyes as a cue that someone is looking at you, or systematic tracking of your motion as a cue that you're being followed (for a similar view, see Krebs & Dawkins, 1984). This will upset some theorists of mindreading because they will argue that it lumps together behavior reading with true mindreading, and even other forms of intersubjectivity. I won't deny that, nor will I deny that there are more sophisticated forms of mindreading that combine cues and stored information to make elaborate inferences and predictions about others' behavior. Indeed, those more elaborate forms of mindreading are what most interest me and are what I study in my own work. However, I think we need to start simply in order to see how these more complex forms of mindreading evolve. And more importantly, it's quite likely that these more complex forms *depend* on simpler forms; we don't lose them; we add to them. Thus, if we want to consider mindreading as a system of interacting parts, some old and some new, some simple and some more complex, we'll need to think about them all—not lumping them together of course, but thinking about how they interact and what design features enable them to do so. After all, things like behavior reading, mutual intersubjective influence, and full-blown conscious reasoning about what another person is thinking all occur; they're not mutually exclusive and may indeed be mixed and matched in various ways in our cognition.

What I'd like to explore here is mindreading as a case study in cognitive evolution. It has all the hallmarks of a complexly organized adaptive system: It likely evolved in steps rather than all at once, and it likely involves the interplay of multiple, specialized mechanisms. Moreover, it almost certainly interacts with mechanisms outside of mindreading proper, such as mechanisms of cultural transmission, language, learning, and so on—often not by accident, but by design. And mindreading is distributed in various forms across species, with (perhaps) its most sophisticated version appearing in humans. The fact that mindreading comes in many forms allows us to make some inferences about how and why it evolved. Add to this the central role

that mindreading seems to play in much human social interaction, and it provides a useful case study for how a putatively modular system can actually add to human flexibility.

---

Much of the literature on mindreading and theory of mind circles around a very specific ability: the ability to compute another's *false* belief about the world and to use this to predict their behavior. The reasons for this are both historical and theoretical (Bloom & German, 2000). Historically, work on theory of mind was jump-started by a paper by Premack and Woodruff entitled "Does the chimpanzee have a theory of mind?" (1978). Here they used the term "theory" in the way psychologist Jean Piaget talked about children as having "theories" of the world, sort of like a scientist does. Premack and Woodruff wanted to know, in essence, whether chimpanzees "theorize" about what's going on in the minds of others. Do they make guesses, for example, about what another individual knows and use these for their own strategic purposes? And how would we know if they are in fact doing this?

The philosopher Daniel Dennett proposed a test that he claimed would show for sure that chimps formed representations of others' mental states, if they could pass it. The logic was that in many cases, it might be possible for an individual to predict another's behavior based on the way the world really *is*—that is, the individual's own representation of it—plus some rules of thumb, without true mindreading. A chimp, for example, might know that other chimps tend to eat bananas or even that they tend to approach bananas and put them in their mouths, without attributing anything as mentalistic as "liking." If I'm a chimp and know where a banana is, I might predict that another chimp will approach and eat it, and this might *look* like I'm predicting his behavior based on a representation of his mental states (he "believes" the banana is in location X and "likes" bananas, so I predict he'll go there and eat it). However, many would argue that this is not true mindreading because there is no representation, even in the form of a minimal placeholder, of another's mental state.

Dennett suggested that if a chimp (or a person) could be shown to predict another individual's behavior based on a *false* belief—one inconsistent with the true state of the world and with one's own knowledge of it—then this had to be because the individual was truly representing a belief, a representation of the world separate from the state of the world itself. This idea spawned a cottage industry of research using various versions of the "false belief task" to test both humans and nonhuman animals—an industry that has produced hundreds of studies to date (Wellman et al., 2001).

There are many versions of the false belief task that test false belief representation and inference in various ways. In the most classic version of the task, sometimes

called the "Sally/Ann task," children are presented with a scenario (e.g., a cartoon strip or puppet show) in which one character hides an object somewhere in a room and leaves. A second character comes and moves the object to a different hiding place. The child is asked where the first person will look for their object when they come back into the room, or where they believe it is (Wimmer & Perner, 1983). In another version, sometimes known as the "Smarties task," a child is shown a box of candy and asked what's inside. They say "candy," and then the experimenter opens the box and pulls out a pencil. He then puts the pencil back in the box and asks the child what another person will say when they are asked what's in the box (Perner et al., 1989).

Chimps, to date, have not been shown to be able to pass the false belief task (Call & Tomasello, 2008). However, some caution is warranted because at first, it was thought that they couldn't attribute belief states to others at all—even so-called true beliefs or knowledge of the way the world really is. In the 1990s, psychologist Daniel Povinelli and colleagues conducted a series of studies showing that chimps ignored whether a person could see them—for example, whether they were blindfolded or had a bucket over their head—when begging for food (Povinelli & Eddy, 1996). Later, however, Brian Hare, Josep Call, and Michael Tomasello showed that it matters a lot *how* you test them. They devised a set of studies in which a subordinate chimp had to decide whether or not to take some food, depending on whether or not a dominant chimp had seen the food and therefore knew/believed it existed. Importantly, tracking what the dominant chimp knows in this case is crucially important for subordinate chimps, because they get beat up if they take food that the dominant chimp knows about. As it turns out, chimps *can* track knowledge in this situation and can even take advantage of cases where *they* have seen the food but know the dominant chimp hasn't: ignorance, which is on the edge of false belief (Hare et al., 2001; see Flombaum & Santos, 2005, for similar results with rhesus monkeys).

There is a similar cautionary tale to be told about children's understanding of false belief. For a long time, using standard false belief tests like the Sally/Ann and Smarties tasks, it was thought that children develop the ability to represent false beliefs around the age of four, give or take, based on a variety of factors. Indeed, by 2001, hundreds of studies had shown this (Wellman et al., 2001). Many people devised versions of the task that could push the age of passing a little younger, but the age barrier was really broken when psychologists Kristine Onishi and Renée Baillargeon devised a way to test children nonverbally. They used a looking-time paradigm in which children viewed a live-action Sally/Ann scenario, at the end of which Sally reached for her object either where she had first hidden it but where it no longer was (false belief), or where the object had been moved when she wasn't looking. Fifteen-month-old babies looked longer when Sally reached for where the object really was—and where the babies knew it to be—suggesting that they were surprised.

In other words, they had generated an implicit prediction that Sally would look where she *thought* the object was, based on a false belief, and were surprised when she didn't (Onishi & Baillargeon, 2005). Evidence for false belief tracking in much younger children has now been found in multiple studies across cultures, with some evidence suggesting false belief tracking as young as seven months (Barrett et al., 2013; Kovács et al., 2010; Southgate et al., 2007; Surian et al., 2007).

What's going on? There is a debate in the literature over two kinds of explanation. One is what philosophers would call the "deflationary" explanation: The nonverbal tasks with babies aren't showing what they claim to show. According to some, babies could just be using a set of behavior-reading heuristics to solve these tasks—for example, rules like "people will look for something where they last saw it" or "people who didn't see something move will look in the wrong place"—without truly attributing beliefs (Low & Wang, 2011; Perner & Ruffman, 2005; Ruffman et al., 2012). Certainly this is possible, though follow-up studies have suggested that these particular two heuristics are not likely to explain the whole of infants' abilities (Knudsen & Liszkowski, 2012; Southgate et al., 2007; Southgate, 2013; Träuble et al., 2010). Another possibility is that babies really are making inferences about others' false beliefs and using them to predict their behavior. On this account, there was something about the original standard tasks that made them hard for babies, just as begging and blindfolds didn't work for chimps, but competing for food did. According to Renée Baillargeon and colleagues, the critical difference is that the task is different when you ask children a direct question (e.g., "Where will Sally look?"). This adds an additional computational burden that is not there when you simply measure children's implicit expectations using looking time (Baillargeon et al., 2010).

Are these explanations truly different? Yes and no. On the account of cognition I'm offering here, everything that the mind does can be described as a set of computations, some of which might be quite complex and some simpler. A heuristic like "people search where they last saw something" is just a rule, albeit a relatively simple one, for generating a prediction about behavior from a prior cue (gaze) that indexes knowledge. On the functional definition of mindreading I've offered here, it *is* mindreading; there is no magical line to be crossed that makes it "really" about mental states, other than that it indexes them in a functionally appropriate way.

That's not to say, however, that all potential mindreading designs are the *same*. There is no doubt that the mindreading capacities of a young infant and a fully-developed adult are not equivalent. In the case of early non-verbal measures of false belief tracking, an interesting and still unresolved question is whether the performance of babies on these tasks and the performance of adults on false belief tasks are being done using the same mechanism.

There are reasons to lean toward the possibility that they are. One is a parsimony-based argument of the "if it walks like a duck" variety: If two individuals of the same species exhibit similar performance on similar tasks it's reasonable to assume, in the absence of further evidence, that the same mechanism is involved. To my knowledge, there exist very few cases where adults and infants have been tested on tasks that are truly the same (for an exception in which infants and adults perform similarly on the same belief-tracking task, see Kovács et al., 2010). Moreover, there are a variety of reasons to think that kids are sensitive to what others think and know, or don't know, well before the age of four (Bloom & German, 2000). If you've interacted a lot with young children, you'll suspect that they are tracking what you're thinking—even if it differs from what they're thinking—well before their fourth birthday. Daniela O'Neill showed that two-year-olds gesture and refer to unreachable toys they want when they know their parents are unaware the toy is there, suggesting belief tracking and, at least, recognition of ignorance (O'Neill, 1996). As developmental psychologist Alan Leslie has pointed out, pretend play (which children engage in as early as two years of age) seems to imply an understanding that people can represent the world differently than it really is (Leslie, 1987). And there are reasons to be suspicious of the idea that representing false beliefs is particularly *hard*, computationally speaking. The computational difficulty of maintaining multiple, mutually inconsistent representations of things is sometimes part of the explanation for why false belief tracking was thought to develop so late, and only in humans. But as I'll describe below, it's not obvious that the computations are necessarily more difficult than the computations required to produce a stereo image by combining information from both retinas, to predict the future locations of a moving object, to embed one phrase within another in a sentence, or to use the same word to refer to two different things. There are lots of things that kids can do well before age four that are seemingly just as computationally complex or more so, and even baby brains contain plenty of neurons to do complex computations.

Just because the computation is neurally plausible, however, doesn't mean that a specialized mechanism isn't necessary to do it. Adding two integers together is computationally easy, for example, but special evolutionary circumstances are necessary in order for organisms to get some kind of fitness advantage from being able to do so (and even humans, it seems, might not be able to do it without special training; Gordon, 2004; Pica et al., 2004; Sarnecka & Carey, 2008). The questions regarding theory of mind abilities, then—including false belief tracking—are questions about the computational design of the underlying mechanisms, how and why these evolved, and how they develop. What species would benefit from tracking others' beliefs, and why? How does this machinery get built, and what evolutionary precursors must be in place for such abilities to evolve? Do theory of mind abilities evolve

in certain contexts like cooperation, competition, or communication, and does this shape their design?

To answer these questions, we're going to have to think more carefully about ecology—including social ecology—and function. What fitness good does it do to make inferences about others' minds? And what kinds of mechanisms can be built to do it?

---

Let's go back to basics and ask a seemingly simple question: What is a mental state such that we can represent it? Consider the following scenario. A zebra is grazing peacefully on the savanna when a lion appears from behind some tall grass. The zebra's object and motion sensors detect that something animate has appeared, and the zebra looks up. Based on some combination of perceptual features (shape, color), the zebra categorizes the object as a lion. And mechanisms in the zebra's brain find the lion's eyes and compute that they are looking straight at the zebra. This kind of inference is where mindreading begins.

First, let's talk about the mechanisms that do this computation. They are variously called eye direction detectors or gaze detectors, and there is evidence for them in many species of animals, from fish to birds to mammals (Baron-Cohen, 1995). Indirectly, we have evidence that they are widespread because of the commonness of a certain form of mimicry called eyespots: structures that resemble eyes (present in taxa as diverse as butterflies and peacocks) that have evolved because of their effects on the eye detection mechanisms of predators or mates. And a variety of behavioral experiments have shown the importance of gaze detection in many taxa. For example, the broken wing display of a ground-nesting bird, the plover, is triggered by looking directly at the bird, and other forms of anti-predator response, like tonic immobility (a form of "playing dead" seen in chickens and other animals) are enhanced in the presence of eyes (Gallup, 1977; Ristau, 1991). Some species might have relatively crude gaze detectors that respond simply to the mere presence of eyes, while others, like us, can actually triangulate what someone else is looking at with a great degree of accuracy—improved by the fact that humans have a white sclera around the pupil, possibly evolved for this reason (Kobayashi & Kohshima, 2001; Leekam et al., 1997).

The reason that gaze detectors evolved and are so taxonomically widespread is not because eyes are inherently interesting. It's because gaze indexes what an individual knows: in particular, what they can see. If I see you, I can know, among other things, *where* you are, *what* you are, even who you are and what you are doing (depending, of course, on my own cognitive abilities). Thus, sensitivity to gaze is an extraordinarily useful thing in a wide variety of contexts, and it opens a slew of pathways up fitness

hills toward ever-more complex ways of keeping track of what others have seen and therefore know.

As I mentioned, many theorists suggest that there is a clear-cut distinction between the mere tracking of perceptual cues, such as gaze, and the genuine representation of mental states, such as knowledge and beliefs (Heyes, 1998; Penn & Povinelli, 2007; Povinelli & Vonk, 2003). On this view, a plover might be just reacting in a reflex-like way to the stimulus of eyes, without attributing a mental state like "knowledge" to the predator. We know that humans sometimes do actually attribute knowledge and beliefs to others because they say so (babies can't, but four-year-olds can). And tasks like the false belief task do show that individual's predictions are based on *some* kind of internal marker or representation of another's belief, because reaching for the "wrong" hiding place can't be predicted just from information that is currently present in the world.

Clearly, there is a difference between behavior that is driven directly by some presently observable cue, like gaze, and behavior that depends in addition on some internally encoded information about a past state of the world, as in the false belief task. But one of the criticisms of looking-time studies of infant false belief tracking is that they could "just" be using a rule, such as "people tend to look for things where they saw them before." Note that deploying this rule requires two things: First, the child must remember where the person looked before (a past state of the world that is no longer perceptually available). Then they must recall this memory and combine it with the rule to predict the other person's behavior. Is there genuinely a difference, from a Shannon information point of view, between "just" using a stored memory combined with a rule and "really" tracking a belief?

Let's consider a simple, imaginary model of how belief/knowledge tracking might evolve (I'm lumping these together temporarily because "knowledge" is the same as "true belief"; we can ask later whether one mechanism or two would be necessary). Imagine that there is a set of neurons in the zebra's head that is activated by the lion's gaze. On a mechanistic level, this is just a description of what a gaze detection system must be: a bunch of cells registering that the zebra is being looked at. Imagine further that as long as the lion is gazing at the zebra, these cells remain active. One can easily envision the usefulness of such a bunch of cells if they are wired to systems that alter the zebra's behavior in a manner that is appropriate to being seen by a lion. For example, activating these neurons could make the zebra more wary, cause it to twitch its muscles in preparation to run, or divert its attention away from the grass it's eating and toward the lion, keeping track of the lion in case flight is necessary. In fact, many prey animals appear sensitive to just how close predators are and what kinds of cues they are giving off (Stankowich & Blumstein, 2005). It's not always best to immediately run when a predator appears, not just because this might trigger the

predator's pursuit response, but also because the predator might have no intention of attacking. It might be full from a recent meal or tired. And predators on the savanna are common enough that running every time you saw one would result at best in a serious waste of energy and at worst in starvation. So gaze detection could activate a wariness mode that then directed the animal's attention and altered its decision thresholds with respect to a variety of cues, both external (distance to predator, distance to plausible escape route, presence of other zebras as cover) and internal to the zebra (hunger, sleepiness).

Now let's imagine a design variant in this gaze-sensitive bank of neurons. Suppose the new variant acts a bit like an egg timer.[1] Its neurons remain active as long as the lion is gazing at the zebra, but then, when the lion stops looking or even disappears from sight, the cells continue to remain active for a period of time, slowly decaying toward zero activation over a period of minutes. This is, then, a "detached" representation in Gärdenfors's (1996) sense, since the neural activity persists after the cues producing it are gone. A slow decay function of this sort could be quite useful, allowing a "wariness buffer": Even if the lion looked away for a moment, the zebra would not immediately relax. Moreover, even if the lion adopted potentially sneaky tactics like going behind some tall grass as it stalked the zebra, the zebra would remain vigilant and ready to run. One could imagine selection favoring this slow decay property of the neurons activated by a predator's gaze, with a speed of decay calibrated over both evolutionary and developmental time by the cost/benefit tradeoffs of relaxing vigilance after having been seen.

Here is an interesting question: What exactly do these cells represent? What information do they carry?

From a Shannon point of view, there are several descriptions of what these cells represent that are approximately, or in some cases exactly, equivalent. Remember that Shannon's definition of information has to do with reducing uncertainty about states of affairs in the world. Structures that carry information are therefore *indexes* of events, properties, or situations. And, as Shannon pointed out, these indexes need not be perfect—indeed, they almost never are. They stand in a statistical relationship to those things they index.

So what do these neurons in the zebra's brain index? For one thing, they index gaze. Note, of course, that they are not a *perfect* index of gaze. The zebra could be being stared at by a lion even though it could not itself see the lion; in that case, the

[1] For those who might be unfamiliar with an egg timer, it's just a countdown clock. When cooking an egg, you set it for a certain number of minutes and let it go. It counts backward to zero, at which point it rings and you know the egg is ready. Here, I'm imagining a timer that starts the moment the lion's gaze is no longer visible and decays slowly to zero.

cells would not be activated (an example of a "type II" error, or a miss). Moreover, the detector itself certainly has some room for error. If sensitive enough, it could, for example, "jump the gun" and fire even if the lion wasn't looking (an example of a "type I" error, or false positive).

What about the decay function? Is that indexing gaze? Well, no—as soon as the lion looks away and the zebra's cells remain active, those cells are obviously no longer indexing being looked at *right now*. But they are indexing something: having *been* looked at. And if the neurons have a decay function, the current strength of activation would be an index of the time since the zebra was last looked at. Such an index could be predictive of a variety of things: probability of attack, perhaps. And, at least until the zebra moves, it indexes what the lion *knows*. For example, even if the lion turns his head—so that he is no longer gazing directly at the zebra—he still knows where the zebra is, at least for a moment. Given that the lion is now not watching, this might be a good time to sneak away.

One thing I'm suggesting with this little scenario is that representational systems for tracking the thoughts of others can be immensely useful—and they needn't be that hard to build, from a computational or evolutionary point of view. Of course, the more complex your mindreading skills are, the more specialized machinery you need. And there must be a fitness benefit in order for such machinery to evolve. The mental states of others are not *intrinsically* useful to know; they need to have a possible effect on your fitness. In the zebra case, a lion's knowledge of your whereabouts is clearly a good thing to track. And in humans, much of our fitness depends on learning from others, coordinating with others, and other forms of complex social interaction. This means that the potential benefits of tracking what is in others' heads and using it to draw inferences, learn, and make decisions are enormous.

Another thing I'm suggesting is that the question of what *really* counts as a mental state representation is a bit of a red herring. What matters, from an informational and evolutionary point of view, is what *function* the representation, index, marker, or computational rule plays. Obviously, no neural pattern has a little label on it that says "this is about a belief." From an informational point of view, the "content" of a representation is not written into it like words in a book (there can exist "iconic" representations that represent, e.g., the visual shapes of words, but these are a special category of representation). The content of a representation, broadly speaking, is a function of what it indexes and how it interacts with other mechanisms within the mind, as in the egg timer's link to a fight-or-flight readiness system. By analogy, representations of "red" in the visual cortex aren't *really* red; they're just neural activity. You couldn't tell they were about redness unless you saw what stimuli caused them, how they were hooked up to other systems, and what cause-and-effect role they played within the system. Similarly, a decay function in a gaze detector is just that: a set of neurons that

decay in activation as a function of time since last activated. They therefore index both having been looked at *and* the lion's awareness of your location. If they are used to decide whether to run or not, this counts as using an index of another's mental state to adjust a behavioral strategy. It is, under the information-based definition I laid out above, mindreading.

Now, what's necessary to pass the false belief task, even in pre-verbal infants, is presumably more than the imaginary egg timer I've described. Where we start to make progress is by thinking about exactly what more would be required in computational terms. First, when the baby sees a person put an object in box A, some representation must be created that links that person and box A. It's probably not merely a decay function, but longer lasting; this can be measured by seeing how long the baby continues to expect the person to look in box A. The baby then sees the object moved to box B *and* also sees that the person didn't witness this. If the baby is then surprised when the person reaches for box B, or is *not* surprised when they reach for box A, the baby must have created a representation that *indexes* the fact that the person thinks it's in A, in the sense of "index" that I've used above. There is really no way to solve the task without an index of this kind. Whatever form it takes, it is an index in that it functionally correlates with what the person saw, and therefore with the person's belief. There may be no further answer to the question of whether the information is "really about" mental states: It indexes them. But, on the account I'm offering here, there may be nothing more to ask. Functionally, you have a system that tracks beliefs.[2]

Thus, I will put aside questions of whether representations and computations are "really about" mental states, unless they are questions about *function*. Does the computation in question *leverage* information about another's internal states, and in what way? On this view, we can see all mindreading mechanisms as part of a spectrum of mechanisms, some simpler and some more complex, that do different things for

[2] Some psychologists might feel that the best evidence that adult humans track false beliefs comes from the fact that they can say so (e.g., Heyes, 1998; Penn & Povinelli, 2007). For example, if you ask adults why the person looked in box A, they might say something like "because they believed the object was in box A." While such self-reports are interesting, it is certainly not a *necessary* feature of representations that we can talk about them. More worrisome than that, there is lots of evidence that peoples' verbal reports often don't reflect the *real* reasons why they decided or inferred something (Nisbett & Wilson, 1977). Much self-report may be what psychologists call "confabulation"—for example, people hardly know the causes of their own behavior better than an outside observer might, and sometimes worse (Wegner, 2002). Thus, there are reasons to take evidence like verbal reports with a grain of salt when trying to uncover the true phylogenetic distribution of mindreading. It's not that language isn't important in theory of mind in humans, but much human mindreading may be nonverbal in form and homologous with that of other species.

different purposes. In keeping with the mental diversity view, this does not mean that all mindreading mechanisms are the same. On the contrary, they are diverse. But it does mean that there is nothing particularly magical about mindreading. It's the result of ordinary mechanisms operating on ordinary information. The question of interest is how such an apparently extraordinary skill—which in humans allows us to do things like watch television, have a conversation, and learn to tie our shoes—is constructed from such ordinary components.

---

A case that illustrates well this functionalist approach to mindreading is the case of corvids—the family of birds that includes crows, ravens, and jays (Clayton et al., 2007). Although these birds may sometimes be annoying (yes, to the crows that particularly enjoy the tree outside my office window, this means you), it turns out that they also perform feats of mindreading, memory, and tool use that have led many to rethink the idea that these things require big brains (Balda & Kamil, 1992; Weir et al., 2002). Corvids do have a large brain-to-body-size ratio, on par with primates and cetaceans, and many argue that brain-to-body-size ratio is more important than absolute size.[3] However, not only are corvids' brains much smaller than human babies'—in the walnut range—they lack a neocortex (though they may use non-homologous structures for similar purposes; Emery, 2006). And their mindreading skills illustrate the importance of thinking about how cognitive abilities are shaped by the specific ecological contexts in which they evolved.

Of particular interest is mindreading in the context of food caching. Many corvids are food-caching birds, meaning that they hide food and recover it later. One species of corvid, Clark's nutcracker, has been shown to be able to remember the individual locations of tens of thousands of cached seeds that it has painstakingly hidden, returning to the exact spots later (Balda & Kamil, 1992). Obviously, these seeds represent a potential bonanza for someone else: If you've done all the work of finding the food, you don't want your cache to be pilfered. Therefore, it's not surprising that corvids are quite sensitive to the possibility of their food being pilfered, and food competition is the context in which their most sophisticated mindreading skills have been observed (similarly, and probably not coincidentally, to chimpanzees).

[3] See Striedter (2005) for an argument that this commonly held assumption might require rethinking. More neurons means more neurons, after all, and while you probably need more neurons to control the operations of a larger body, there is no principled reason to expect a simple mapping function between brain and body size.

In the field, biologist Bernd Heinrich observed that ravens are reluctant to cache food in the presence of conspecifics. They will wait to hide it until the other leaves or, even more interestingly, until the observer looks away or their view is obscured by a barrier (Heinrich & Pepper, 1998). This was confirmed experimentally in a series of studies on scrub jays by Nicola Clayton, Nathan Emery, and Joanna Dally, which showed that jays make strategic use of barriers to hide food when a conspecific's view is blocked, and that they adjust their caching pattern according to a variety of factors, including the relative dominance of the observer (again, reminiscent of chimps). In addition, jays will re-cache food that they know others have seen them hiding. They actually distinguish between those who saw them hiding the food and those who didn't, re-caching more in situations where they are confronted with a knowledgeable versus an ignorant observer. And in the most interesting twist of all, jays who have stolen food from others' caches in the past are more likely to re-cache their own food when being observed—a case of "it takes one to know one." This implicates learning, of course, and a role for personal experience in shaping the behavior—as well as, possibly, perspective-taking or putting oneself in another's shoes (Clayton et al., 2007; Dally et al., 2005, 2006; Emery & Clayton, 2001).

Of course, there are various possibilities for the design of the underlying mechanisms here (see Emery and Clayton, 2008, for a proposal). However, the behavior clearly qualifies as mindreading in the sense that I've defined it and goes beyond mere contingent reaction to directly observable cues. For example, distinguishing between individuals who have and haven't seen you cache a piece of food means that you must remember those individuals separately and associate a perceptual event (e.g., gaze) with one of those individuals—a gaze event that is now in the past and therefore indexes stored knowledge in the other individual. Not only that, Clayton and colleagues' results suggest that this knowledge is associated with a *particular* cache, suggesting an even more sophisticated belief representation system (individual X knows that I hid this piece of food here, but didn't see me hiding that piece of food over there). Add in other features, such as what Clayton et al. call "experience projection"—using one's own experience as a thief to modify one's expectations about others—and you seem to have a fairly complex system, composed of multiple components and designed to track others' knowledge states specifically in the context of food theft.[4]

[4] To the extent that the skills seen in humans, scrub jays, and dogs (discussed below) are similar, these presumably represent cases of convergent evolution of similar skills in distantly related taxa. They thus present a nice case study for the comparative method, asking what are the similar ecologies and evolutionary histories of these species that led to (at least partly) similar forms of mindreading evolving in each.

There is room, of course, for skepticism as to whether jays are really that smart, and as everywhere else, the particular details as I have them here could prove to be wrong. But assuming that Clayton, Dally, and Emery's account of scrub jay mindreading is right, this looks like a system of representational formats and decision rules that evolved to solve a particular adaptive problem: avoiding food theft. It remains to be seen if scrub jays' mindreading abilities extend outside this domain, but it seems quite likely that scrub jays can't represent just *any* belief. Whereas we, for example, can form representations like "Rick believes in Jesus," scrub jays probably can't. This is probably, as we'll discuss below, at least in part because they can represent neither Rick (as an individual human with the identity "Rick") nor Jesus (as an individual deity, "the son of God"). But they *do* seem to be able to form representations that index beliefs, something like "Scrub jay number 24 saw me hiding a seed at location number 83." Thus, from a comparative perspective, tracking beliefs is not just a single thing—there are different versions of tracking beliefs, tuned to solve different ecological problems. This is quite different than the idea that belief representations are a kind of general cognitive milestone reached only by humans. Instead, different species may track many different kinds of mental states for many different purposes relating to their particular modes of living. A question that remains, however, is this: Why are humans so much more flexible—or at least, so it seems—in the beliefs that we can represent? Why is the *scope* of beliefs we can represent so much larger, and via what mechanisms is this accomplished?

---

Before we turn to ourselves, let's look at one more example: dogs. Recently, a burgeoning subfield of dog studies has developed in comparative psychology, specifically examining dog skills of social cognition and mindreading (Miklósi, 2007). And, as with corvids, this work is revealing that sophisticated mindreading skills have evolved in more than one lineage and are probably due to more than *just* having large brains. In the case of corvids and possibly chimpanzees, it appears that the primary selection pressure for mindreading came from competition, and specifically competition for food. Canids, however—dogs and wolves—are able to mindread in a different context, and for a different purpose: cooperation, and in particular, cooperative communication. And perhaps you won't be surprised to learn that they appear to be particularly good at communicating about food (yes, food again).

Dogs, of course, evolved from wolves via a process of human domestication (Wayne & Ostrander, 2007). Many authors have suggested that it is not a coincidence that our species became mutualistically intertwined: Humans found a lot that they could relate to in wolves. Wolves have two features that make them similar to

humans: They are highly social, and they hunt. And these things go together. Wolves are *cooperative* hunters, just as humans are, and have been so for a long time.

Cooperation, defined biologically, implies collaborating for mutual benefit. Mutualism is the word that biologists use to describe win-win situations. For example, nitrogen-fixing bacteria live in the roots of plants and thereby gain protection and nutrients. The bacteria, in turn, fix nitrogen, providing the host plants with a nutrient crucial for life, one that plants can't make themselves. Altruism refers to cases where one individual pays a fitness cost to provide a fitness benefit to another (e.g., giving money to someone on the street would represent a true case of altruism). Biologists suspect that true altruism—behavior that leads to net fitness loss over the altruist's lifetime—is probably rare, because systematic fitness loss can't be selected for. But when altruists help each other, it's possible to get a win-win situation if the benefit each receives is greater in the long run than the cost paid. This is how most cooperation works: It is, ultimately, mutualistic. The various parties do better by cooperating than they would if they didn't.

The literature on cooperation is huge and notoriously contentious, and I am glossing over many details here; there are nuances and exceptions to my rough summary (see Hammerstein, 2003, for a sample of contemporary views). But mutualistic altruism is a good description of cooperative hunting, because hunting is only cooperative if at some point, individuals pay some kind of fitness cost—opportunity cost, for example—in order for the hunt to be successful. When the hunt *is* successful, all individuals are, on average, better off by the sharing of the spoils. Importantly, some hunting may *appear* to be cooperative to an outside observer, as when many animals swarm to try to catch a prey animal. However, if each is only trying on his own to get the prey and never does something altruistic in order to help another succeed, then it's not genuinely cooperative (this is sometimes called facultative cooperation or byproduct mutualism, whereby each individual acts in his own interests but incidentally helps others by, e.g., flushing out game and making it easier for them to catch). For this reason, there is some controversy about whether and when chimpanzee hunting is genuinely cooperative (Boesch, 1994; Mitani & Watts, 2001; Tomasello et al., 2005). For wolves and other canids (as well as hyenas and felids), however, the present consensus seems to be that they do cooperate to take down game (Macdonald, 1983).

Communication is not strictly necessary for cooperation to occur, but it helps. Moreover, most communicative situations, if they are evolutionarily stable, involve a form of cooperation, even if it doesn't seem that way. For example, if a thief says "give me your wallet or I'll stab you," most of us wouldn't describe the situation as cooperative, but in fact, the communicative element is. Rather than just stabbing you, the thief has bothered to tell you something, presumably because ultimately he'll be

better off than if he didn't (e.g., he reduces the risk of you fighting back). And if you *believe* him—if the communication is successful—then you'll give him your wallet, paying a cost to avoid a fitness-worse outcome. Given the situation, it's win-win.

What does this have to do with hunting and with mindreading? The idea is that wolves, or any other cooperative hunter, would benefit from being able to infer the plans of their fellow hunters and coordinate appropriately. There would also be advantages to communicatively strategizing: "I'll go this way, and you sneak around that way and ambush him." Of course, this kind of strategizing only makes sense if you take what individuals are saying as an index of their true internal states. Most communication, in fact, involves mindreading as I've defined it, since making a signal provides a cue to your intentions, and using the signal involves treating it as such (Krebs & Dawkins, 1984; Maynard Smith, 1991; Sperber & Wilson, 1995). Even if you suspect someone is lying, you're mindreading (or attempting to), since a lie involves a mismatch between an utterance ("I intend to do X") and a true internal state (I don't intend to do X).

There is good evidence that dogs are able to mindread communicatively. What started the experimental ball rolling was a set of studies by Brian Hare and colleagues in which they compared the abilities of dogs, wolves, and chimps to understand communicative pointing by humans. This involved a setup in which there were several possible hiding places for food, which the animal knew—but they didn't know *where* food was hidden in a given case. In each trial of the experiment, a person would simply point to a hiding place, and the question was whether the subject—dog, wolf, or chimp—would use the pointing as a cue, following it to find the hidden food. Results suggested that dog puppies did follow the cue above chance, but neither wolf puppies nor adult chimps did (Hare et al., 2002; Hare & Tomasello, 2005).

Hare and colleagues interpreted these results in light of domestication. After all, for the experiment to work, the animal must in essence believe that a communicative cue from a human will help it to find food. Whether or not chimps hunt cooperatively in the wild, they don't appear to do much cooperative communication, and certainly never would have been selected to attend to *human* communicative signals. Wolves might communicate among themselves, but again, why trust a person or listen to them? Dogs, on the other hand, have been selected specifically to cooperate with humans, and even to communicate with them. For example, in the context of hunting, dogs will both give information to humans (e.g., pointing at prey) and understand and obey human commands.

In the intervening years, many more studies have been done, and the picture has become more complicated. Again, we might expect context to matter here. Wolves might not, in the absence of special training, pay attention to a human communicative cue—but they might be able to mindread cooperatively among themselves.

Consistent with this, there is evidence that wolves can do better on tasks involving communication than Hare et al.'s original results implied, and that the difference has to do in part with the effects of developmental environment on attention to human cues (Miklósi et al., 2003; Udell et al., 2008). In addition, the basic ability in dogs has been replicated and expanded to include, for example, taking into account what humans know in a communicative context. One study, for example, showed that a dog will whine more when a person doesn't know where a hidden treat is located, as if trying to tip them off that it's there, reminiscent of O'Neill's finding with children (Virányi et al., 2006). While the exact nature and boundaries of the ability are still being investigated, a growing number of studies suggest that dogs are sophisticated mindreaders, and they do it cooperatively. If you have a dog, you're probably not surprised to hear this.

The case of dog mindreading nicely complements the case of corvid mindreading in a several ways. As in corvid food competition, the dog results suggest that mindreading may, at least in some cases, be tied to specific ecological and interactional contexts. Where an animal might fail to mindread in one context, it might succeed in another. Second, it shows that the nature of mindreading skills is, at least to some extent, fit to social problems that a species faces. Jays face a serious threat of food theft and so are sensitive to what potential thieves know. Dogs get most of their food either from people directly or in cooperation with people, and so it makes good sense for them to be sensitive to cases where humans are trying to help them. And the contexts of mindreading needn't be purely social, in the sense of dealing with conspecifics: Predators may be selected to be good at mindreading prey and vice-versa (Barrett, 2005a).

The case of humans, however, is even more intriguing. Unlike corvids, canids, and chimps, humans can mindread in both competitive and cooperative contexts, and can engage in trains of reasoning and inference about others' mental states that, presumably, no other animal can entertain. "John doesn't think that Sarah knows we're going to the wedding," for example, seems far beyond what any other species can do, both in terms of content and complexity. What machinery allows us to do this kind of thing, and how did it evolve?

---

Humans mindread for many diverse purposes, presumably much more so than jays, wolves, dogs, or chimps. Whereas we can make inferences about what others are thinking in contexts as varied as game shows, job interviews, and military battles, the skills of these other species, while impressive in their own right, seem more limited to specific contexts like food theft or cooperative hunting. They may be more

limited in the types of mental states they are able to represent, as well. Dogs, for example, might know when their humans are angry at them, but probably not when they are embarrassed.

Here again, it makes sense to think of contextually oriented mindreading abilities as islands of competence (Frankenhuis & Barrett, 2013). On this view, the ability to represent beliefs is not just a thing that an organism has or doesn't have. Nested within the ocean of beliefs that it *might* represent are the ones that it actually does or can: particular kinds of beliefs in particular contexts. This contrasts with a view in the theory of mind literature that treats belief representation as a kind of black box into which any belief can be slotted: Either you have the box or you don't. It is this view that makes Premack and Woodruff's question "Does the chimpanzee have a theory of mind?" one that can be answered with a "yes" or a "no."

Even in human babies, of course, it's likely that early developing belief-tracking abilities are contextual, representing a subset of what adults can do. We know, for example, that babies can track others' beliefs about the locations of objects, and some other beliefs as well, such as those about the relationship between objects' appearances and properties that can't be seen, including what's inside them or what sounds they can make (Barrett et al., 2013; Song & Baillargeon, 2008). However, it's virtually certain that these islands expand with age and knowledge. Babies don't yet know what Democrats and Republicans are, so it's not surprising if they can't yet represent beliefs *about* them. And the contexts in which mindreading occurs presumably expand as well; people on dates and politicians on the floor of the Senate make mental state inferences that babies never could. By adulthood, the human capacity for mindreading appears much more broad and flexible than even closely related species like chimps.[5]

Why have human mindreading skills expanded compared to those of our nearest relatives, in both the contexts in which they can be applied and the contents that can be represented? We are as yet only beginning to understand the alterations in brain

[5] Let me be clear that by adopting a broad view of what mindreading is, I'm not claiming that all mindreading abilities are the *same*, or denying that humans are likely to have additional mechanisms that are not present in other species. Indeed, that's the point: Mindreading is a good example of an evolutionary mosaic of mechanisms, some new and some old. The complex and flexible human skill of mindreading is likely to be the result of many specialized mechanisms interacting, and it's even the case that not all mindreading events in humans need to recruit the same mechanisms. Some mindreading, for example, is unconscious cue reading, and some involves explicit, conscious reasoning, perhaps represented in natural language. Some mindreading mechanisms are likely to be widely distributed across taxa and some narrowly. Some are likely to be homologous, like gaze following in mammals, and some analogous (convergently evolved), like knowledge tracking in corvids, dogs, and humans.

mechanisms that enable mindreading abilities in humans, and how these alterations add to the skills that were present in the common ancestors we share with other primates. Neuroscientist Rebecca Saxe, for example, suggests that while other primates can represent dyadic relations between individuals and objects—including mentalistic attitudes toward those objects, such as seeing and wanting—only humans represent triadic relations such as "she saw me hide the food" (Saxe, 2006; scrub jays seem to do this, but perhaps they represent a convergently evolved case). Psychologist Michael Tomasello and colleagues argue that it is a specific *kind* of triadic awareness that distinguishes human intersubjectivity—in particular, two individuals jointly collaborating on a shared goal and being mutually aware of the collaboration (Tomasello et al., 2005).

There are a growing number of brain mapping studies looking at the neuroanatomical basis of mindreading that show, perhaps not surprisingly, that many functionally specialized regions are recruited in mindreading tasks (as we will discuss in more detail later, collaboration between specialized systems is probably the norm rather than the exception in cognition). Some regions, such as the medial frontal cortex (MFC) and the superior temporal sulcus (STS), are particularly implicated in mindreading tasks, such as false belief tasks. While the exact functions of these regions are still matters of debate—and there is evidence for further functional subdivisions within them—neuroscientists David Amodio, Chris Frith, and Uta Frith suggest that the MFC is responsible for, among other things, decoupling mental state representations from representations of physical actions, and the STS is involved in agency detection and prediction (Amodio & Frith, 2006; Frith, 2007; Frith & Frith, 2003). Whether any of these regions has been selectively modified specifically for mindreading is not clear, though it is certainly possible given the unique elaboration of the frontal cortex in humans (Striedter, 2005). Saxe suggests that one region, the dorsal medial prefrontal cortex, may be specifically dedicated to computing triadic relations and could be uniquely modified in humans, though this awaits confirmation (Saxe, 2006).

It is an interesting feature of human mindreading that our capacities for belief attribution do seem to be black box-like in the sense that we can represent virtually any belief formed from concepts that we understand. We can represent someone's belief in Jesus or quarks, if *we* are able to represent Jesus or quarks. And it is not necessarily only the belief-representing aspect of theory of mind that has been altered in humans compared to other primates. For example, the fact that we are able to cooperate at much larger scales than any other known primate, or mammal for that matter, suggests that mechanisms such as empathy toward our fellow humans might have been altered as well. What selective factors have favored the changes we see in humans, and what exactly *are* those changes at a cognitive level?

As I'll discuss in greater detail later, I think it's unlikely that there is a single prime mover or fitness benefit underlying the evolution of human mindreading skills. The reason is that mindreading is used in a variety of fitness-relevant contexts. For example, in humans, mindreading appears to be useful for *both* competition and cooperation. Byrne and Whiten's idea of "Machiavellian intelligence" and Nicholas Humphrey's notion of "social chess" capture this nicely. In politics writ large, including the politics of groups and the politics of personal relations, mindreading is useful for both forming bonds and maintaining friendships, and for outwitting one's opponents (Humphrey, 1976; Byrne & Whiten, 1988). In strategic situations such as chess, and game-theoretic situations more generally, being able to guess what your opponent knows and is planning is a pathway to success. This is consistent with the important role that mindreading, broadly construed, plays in our sense of morality: We often judge the reasons for peoples' acts, such as their motivations, and use these to make moral judgments (Gray et al., 2012; Mikhail, 2011; Young et al., 2007; Young & Saxe, 2009). But as we've seen, mindreading is used in other contexts as well, contexts that, not coincidentally, humans are very good at and that play an important role in our fitness. One that we've discussed at length is social learning: Being able to guess what another person is trying to do or trying to show you can give you a huge boost. And in language, mindreading is useful not only for *learning* language—by solving problems of opacity, as in the gavagai scenario—but also for *using* language. Many communication theorists, such as Paul Grice, Dan Sperber, and Deirdre Wilson, argue that covert acts of mindreading—mindreading that we're not necessarily aware of doing—underlie almost every communicative event (Grice, 1989; Sperber & Wilson, 1995).[6]

If this is right, then it might not make sense to look for a single factor or mechanism that makes human mindreading unique. There may have been multiple changes in brain design, and these may involve changes in interactions among multiple systems, including tweaking preexisting systems so that they can interact in ways that were not possible before: small enabling changes that yield substantial cognitive differences.

One interesting possibility along these lines was proposed by developmental psychologist Alan Leslie (1987). Leslie's proposal was made in the context of a long-standing debate about the underlying cognitive capacity that enables belief tracking. Some have proposed that the ability to represent others' mental states is actually just a particular instance of a more general ability: metarepresentation, or

[6] See Duranti, 1988, and Rosaldo, 1982, for arguments that this is not necessarily true of all communicative acts. Opinions might differ depending on whether one takes a broad or narrow view of what counts as mindreading.

the ability to represent representations (Perner, 1991). What the concept of metarepresentation implies is a form of recursion, in which one process is applied to the output of the same process. For example, imagine taking a picture of a cow. You now have a representation of a cow. If you then take a picture of that picture of a cow, you now have a metarepresentation of a cow: a representation of a representation of a cow.

Psychologist Josef Perner suggested that when children solve false belief tasks or otherwise represent others' beliefs, they are engaging in a form of metarepresentation. After all, if I believe that you believe your car is parked outside your house, then I have created a belief about a belief. Perner suggested that doing this with respect to mental states is part of a more general ability to represent representations; for example, when we understand that a map is a representation of physical space, we are using this ability. There is evidence both consistent with and inconsistent with this view. For example, many studies have shown that children with autism are impaired on false belief tasks compared to other tasks (Baron-Cohen et al., 1985). One study found that this impairment did not extend to false photographs—pictures of states of affairs that have since changed, meaning that the photos are out of sync with reality (false) (Leekam & Perner, 1991). If so, this suggests that not all representations are necessarily equally easy to represent, or represented using the same mechanisms. Other findings suggest that some representations (pictures) are generally easier for kids to understand than others (scale models) (DeLoache, 1991). However, there remains a debate about whether different metarepresentation tasks are computationally equivalent, and what this says about the domain-specificity of belief tracking (Leekam et al., 2008).

Against the backdrop of ideas about representing representations, Alan Leslie suggested what is, in some ways, a simpler explanation for the representational formats underlying belief tracking (though it's worth noting that in psychology, one person's "simple" is often another person's "complicated"). Rather than true recursion in which one process (representation) is applied to the output of the same process (representation), Leslie suggested that belief representations could be formed using what are, in essence, representational tags. He noted that representations of others' beliefs, and some other kinds of mental states as well, could be thought of as "propositional attitudes," or attitudes toward states of affairs, and that that these propositional attitudes could be thought of as having three parts. There is an *agent*, the person whose beliefs we are representing (e.g., Sarah). There is a *state of affairs*, some representation of the world that the agent believes (e.g., that dinosaurs and humans coexisted). And then there is a kind of tag that links these two representations, which is the *attitude*, a mental state that Sarah has toward this state of affairs (e.g., [Sarah]—[believes that]—[dinosaurs and humans coexisted]). Leslie called this three-part representational format an M-representation. It recalls the idea of metarepresentation because it is, in a

sense, a representation containing other representations within it; you could call it a compound representation. But Leslie called it an M-representation to distinguish it from the idea of *recursive* representation, and to point out that it has a more specific representational format than just one representation embedded within another (Leslie, 1987).

Now, Leslie's proposal might or might not turn out to be a correct description of how belief representations are formatted in the mind. I suspect that something like it is likely to be right, but my point is less to defend the particular empirical proposal than to point out some of its interesting features and virtues from an evolutionary point of view.

Perhaps the most interesting feature of M-representations is that they make heavy use of evolutionarily preexisting parts—something that is common in evolution in general and probably common in brain evolution as well. Notice, for example, that part of the representation [Sarah]—[believes that]—[dinosaurs and humans coexisted] is the representational component [Sarah], which is *itself* a representation. As we've seen, many species, including humans, are able to represent individual conspecifics (i.e., tag them with a specific identity, recognize them as *that* individual when they see them, store information about past interactions with them, and so on). The ability to represent an agent with an individual identity does not require theory of mind and could (potentially) predate it evolutionarily. Thus, part of what was needed in order to form M-representations already existed (perhaps) prior to the evolution of whatever addition enabled M-representations to be formed.

In addition, [dinosaurs and humans coexisted] is itself a representation. As it happens, it's unlikely that chimps or any nonhuman animal could entertain such a representation. But chimps and indeed virtually all animals with brains are able to form representations of *some* states of affairs. For example, in previous chapters, we saw that the visual system can assemble representations whose form is in some way computationally analogous to *the apple is on the table* or *it's dark*. Chimps and young babies can presumably form such representations. These could—if an evolutionary change occurred enabling M-representations to be formed—become part of M-representations, as in [Sarah]—[believes that]—[the apple is on the table]. Indeed, Hare et al.'s results with chimps suggest that they *can* form representations such as [dominant chimp]—[believes that]—[the apple is behind the barrier].

All that would be missing, then, would be some way of linking up representations of agents and representations of states of affairs in such a way that the agents could have *attitudes* toward those states of affairs. This would, of course, involve some kind of neural rewiring—an appropriate connection forming between the part of our brain that stores representations of agents, and the part or parts of our brains that represent states of affairs (probably many). But it couldn't just be any connection; it

would have to carry information *about* the agent's attitude toward the state of affairs. And this could entail a system of tags corresponding to things like believing, seeing, wanting, hating, liking, knowing, and the like.

In case this seems too complicated to have arisen in a single evolutionary leap, think back to the example of the "egg timer" neurons in the zebra. The initial evolutionary change (a mutation, gene regulatory change, or environmental change—see chapter 12) could be modest and would just entail having a resulting bit of neural tissue with some felicitous properties, just as the egg timer neurons did. At first, for example, there needn't have been multiple mental state tags. Perhaps the first tag was just something like "sees," "knows," or "is aware of," as in the scrub jay: [individual X]—[sees]—[me hiding food]. Once such a system initially appears, even if crude, a pathway up an adaptive landscape is opened such that tags can be duplicated, altered, take new functions, and so on.

Later on, we'll get into more detail about what kind of evolutionary and developmental scenarios are necessary to explain such things, but for now, note that what looked like a quite complicated and structured representational system could, in principle, be initiated through some modest changes to systems already in place. And as we'll see, there is evidence that things like "rewiring" events—putting a little neural connection where there wasn't one before—can provide grist for the mill of selection, making use of preexisting materials in new ways that can be taken hold of and shaped by selection (Anderson, 2010; Barrett, 2012; Dehaene, 2009).

Additionally, this scenario suggests that the scope of things that could be represented by a species' theory of mind could have a lot to do with representational capacities that are *not*, strictly speaking, theory of mind. On Leslie's account, for example, M-representations are formed by combining representations produced by other specialized systems. This has the testable implication that the compound representations thus formed will be constrained by the design features of the systems that generate each representational component. For example, while many species, from scrub jays to baboons to humans, are able to represent and remember individual conspecifics, there are presumably large differences in the richness and content of those representations. Monkeys like baboons and capuchins, for example, represent the social rank and kinship relations of other individuals in their group, and aspects of their history of dyadic interaction as well (Cheney & Seyfarth 2007; Perry & Manson, 2008). Scrub jays remember who has seen them, probably attaching this to a representation of some physical features of the individual, and perhaps a few other things about their past history of interaction (Dally et al., 2006). The contents of our representation of [Sarah], on the other hand—or at least the information stores to which [Sarah] is linked—can be vast. But as in other species, a mindreading event regarding Sarah would begin with the activation of our representation of Sarah, starting either by

observing Sarah herself and having our face and body recognition areas activated, or having the representation [Sarah] be activated by, for example, mention of her name or recall of an episode in which we saw her (see Saxe, 2006, and Amodio & Frith, 2006, for a review of subsystems activated during mindreading events).

The same goes with the scope of states of affairs we can represent: not just [I am hiding this seed] but [dinosaurs and humans coexisted], [the earth is 4,000 years old], and much, much more. If the account here is correct, then part of the reason that we can represent [Sarah]—[believes that]—[dinosaurs and humans coexisted] and chimps can't is because of the additional representational systems that theory of mind interfaces with. And if it's true that M-representations are formed using something like mental state tags, it's quite likely that the set of tags we can draw from is much larger than in other species. Jays and monkeys might have [sees], [knows], [wants]. Humans have things like [doubts], [is contemptuous of], and probably many others.

One question that arises is whether and how human mindreading abilities have (or perhaps haven't) been shaped by the specific contextual uses to which they are put: for example, in linguistic communication, in social learning, in judgments about the goodness or badness of others' acts, and so on. With the advent of sophisticated brain mapping technologies such as fMRI, it's becoming possible to see how particular tasks, like mindreading, recruit different areas and abilities in the brain. For example, there is evidence that areas implicated in mindreading—such as areas computing intentions and empathy—are activated when subjects are asked to make moral judgments (Young et al., 2007) and when they play cooperative economic games in the laboratory (McCabe et al., 2001). Are there designed interfaces between diverse systems that make use of the resources of mindreading and that collaborate to produce judgments and decisions? My guess is that the answer is yes, but we're just beginning to find out.

Another question is about how all this develops. As I've mentioned, when you propose the existence of a specialized mechanism or system, people assume that it's innate. And indeed, in some sense, it must be: There is a reason why one species can do a form of mindreading that another can't. But the ability needn't be innate in all of the intuitive ways in which we frequently use the word. For example, whatever theory of mind might be, it doesn't necessarily have to come fully formed like a machine in a box and simply unpacked and plugged in during infancy. Indeed, it's almost certainly *not* that way. Not only is individual experience likely to matter in shaping mindreading abilities, we might expect experience and learning to be used as a matter of evolved design. Not everyone's mindreading abilities are likely to be identical, appear at the same time, or be deployed in the same way. Even the *contents* of what we are able to represent are almost

certainly heavily dependent on experience—again, by design. If I know that Sarah is a member of the Republican political party in the U.S., I may be able to make guesses about what she believes or how she intends to vote on a specific issue; but this clearly depends on learned background knowledge about Republicans. It's even possible that some people might have different mental state tags/concepts than others. A classic example would be the German concept of schadenfreude, or delight in the suffering of others: something most of us can easily understand but that might require experience to build into our representational repertoire. And anthropologists have suggested that mental state concepts, such as emotion concepts, differ subtly across cultures in what they imply and how they are used to predict and interpret behavior (Lutz, 1988). None of this is at odds with mindreading as an evolved capacity.

If, as I've been suggesting, functionally organized systems of interacting parts don't arise merely by chance or through a general-purpose walk through developmental possibility space, then how *do* they arise? How do we conceptualize the development of design that is at the same time evolved and shaped by natural selection, but also developed and shaped by experience? Here as elsewhere in cognition, I think we can build a theory that encompasses both, not as separate ingredients but as ingredients that exist because of each other. To do so, however, we'll have to go back to basics and rethink some deeply held ideas about how development works.

PART III

# Development

# 6

# DEVELOPMENT

In animals, development is the process whereby organismic design is reincarnated each generation from tiny fertilized eggs. It's the process whereby bats come to have bat wings instead of mouse legs, cows come to have cow noses instead of human noses, and giraffes come to have giraffe necks instead of gorilla necks. It's also the process whereby chickens come to think in the way they do about sitting on eggs, how dogs come to think about urinating on trees and sniffing other dogs, and how humans come to think what they think as they speak, dance, cook, argue, and contemplate. But development is not *merely* the process whereby these things come to happen, as if by chance: It exists *in order* to make these things happen. Depending on how you interpret me, this might seem a shocking claim. Am I saying that natural selection selected *specifically* for dancing and cooking? What I will be arguing is that these are particular tokens of *types* of abilities that we are evolved to be able to develop. Exploring this question—what development is for, why it exists and has the properties is does—is my purpose in this chapter and the next.

Here is something development is not: It is not the unfolding of a tiny, preformed homunculus, or a process where a mysterious agent, a builder, consults a picture of what the developed organism will look like and then just assembles elements to look like the picture. Here is something else that development is not: a general-purpose learning process where the organism starts with no particular plan but just tries out different ways to be as it grows and then picks the one that works out best. At minimum, we know that these views of development are wrong because organisms that start with the "same" picture or blueprint, such as genetic clones, can develop differently depending on the environments they are placed in, and organisms from different species can develop quite differently even when placed in the same environment.

You might have noticed that so far in this book I have distanced myself from the concept of innateness as a universal explanation for the properties of mental adaptations. This is not because I don't think that the idea of innateness can be, and sometimes is, used sensibly. It's because I think that the term "innate," as it is often used, conjures up phenomena that are biologically rare: structures that are fully formed at

birth, every detail filled in, and that never change again. The term "innate" is used by people on both sides of the nature/nurture debate—which is framed, after all, in terms of innateness—from hard-core "nativists" to hard-core "empiricists" (see, e.g., Spencer et al., 2009, and papers in the same volume). There is a large technical literature on the various shades of meaning of the term (Griffiths, 2002; Samuels, 2002, 2004). While different people use the word in different ways, its most common usages tend to refer to whatever is left over when you subtract what develops after birth, or what is learned, or what is not influenced or determined by genes. Samuels defines innate traits as those "not acquired via any psychological process or mechanism" (Samuels, 2002, p. 234). Spelke and Kinzler state, quite simply, that "innate means not learned" (Spelke & Kinzler, 2009, p. 96). Karmiloff-Smith, while urging to put the nativism/empiricism divide behind us, still relies on a distinction between "prespecification" and "plasticity for learning" (Karmiloff-Smith, 2009, p. 101). And other construals of innateness exist, equating it, for example, with canalization (Ariew, 1999).

I tend to agree with Griffiths's (2002) argument that the concept of innateness is, in his words, "irretrievably confused." Yet there is something we need to retain. All accounts of development seem to converge on the idea that some kind of evolved structure, some kind of stuff—including but not limited to DNA—is necessary in order for development to occur. The problem is that many current models of what "innate structure" might be are incredibly impoverished, in that they view this structure as inherently passive.

On one view, what is innate is like a set of bowls or containers into which the richness of experience is poured: The bowls themselves don't really do anything. The only active parts of the developmental process are the learning algorithms, which—because they are simply transferring experience into the bowls—don't have much inherent content, or even assumptions about what they are processing, at all. To the extent that there are any assumptions, those are sometimes held to come in the form of a separate, innately frozen template or picture, or innate representation, which the learning algorithm consults in doing its job (for a discussion see Elman et al., 1996).[1] Again, this kind of template is inherently passive, like a gelatin mold. This points to another commonly held view about innateness: that the innate part that is specified by natural selection, be it a template, container, or what have you, provides constraints to an otherwise open-ended, malleable developmental system. One can hardly think of anything more passive than a constraint, the bolt that keeps the active R2D2 of development from wandering off in unwanted directions.

[1] Recall also, from the introduction, Panksepp & Panksepp's evocative formulation of "unique and detailed epistemological engravings of sociobiological strategies" (Panksepp & Panksepp, 2000, p. 110).

I don't mean to imply that *some* developmental processes might not be like this, involving an inherently passive and static template that is consulted or used by an inherently active, yet content-free, developmental process such as learning. But we know that the biological development of other structures of the body, like limbs and organs, is not like this at all (Carroll et al., 2005; Coen, 1999; Maynard Smith, 1998; Shubin, 2008). To build an arm, you don't have an innate representation or picture of arm properties coupled with a general-purpose developmental algorithm that consults this picture in combination with input from the environment to build it. Instead, the entire process is active; there is no separation of the innate and non-innate parts. Developmental outcomes emerge from the properties of the developmental processes themselves, interacting with the experiences or environments, both internal and external, that occur at every stage of the process. This means that whatever is built in cannot be, or is not usually, separate from the developmental processes or procedures but must be implicit in them, in their structure, such that the proper outcomes emerge. Moreover—and though one can, as I mentioned in chapter 1, see these things as two sides of the same coin—what is necessary is not so much innate constraints to prevent the bad things from happening, but generative structures that make anything happen at all, and that make it be the right thing. This is a much different picture than the binary model of an innate kernel plus learning: There are not two kinds of things, the innate and the non-innate, but only one, the developmental process itself.

As it turns out, historically, it's taken us a while to learn this lesson about the active nature of development. Discoveries about the biology and genetics of how development occurs, most recently in the fields of developmental genetics and evolutionary developmental biology or evo-devo, have generally been taken by the scientific community as a series of "surprises," from the surprise that the phenotypic richness of developed organisms appears to be far underspecified by the informational richness of the genes they carry (what biologist Paul Erhlich has called the "gene shortage," 2000), to the surprise that the genetic system is not merely a set of passive blueprints for proteins, as was thought by Watson and Crick, but more like an active computational machine of switches and pathways that not only assembles proteins but controls them and a host of other biochemical structures as well, including itself. This has been known in developmental biology for some time, but many psychologists continue to think about gene-phenotype mappings as one-to-one, leading to gene shortage-type arguments (i.e., that the brain must contain very few specialized adaptations because there aren't enough genes for very many; see Marcus, 2004).

With all due respect to the benefits of hindsight, I sometimes feel that it is foolish of us to continue to be surprised by each new discovery of the active, interactive

nature of development, and by the discovery that the source of the built-in information is not necessarily where we thought it was, in tiny pictures of developmental outcomes specified in the genes, but rather in the *way* genes interact with each other, with the biochemical soup in which they are embedded, and with the world. But it seems even more remarkable that the obvious has not yet filtered into much of developmental psychology: that we should regard the systems that build our brains and minds the same way as every other developmental system that we know of, at least in one particular sense. Like everything else in biology, these are systems that have been prepared and shaped by the evolutionary process to *cause certain outcomes to emerge* by means that will almost certainly appear ingenious, bizarre, and unexpected to us when we figure them out, but that will almost certainly not be simple.

So what is development like, then? I've made some statements about what it is not—neither preformed, innately specified outcomes nor outcomes that "merely" emerge without design—but is it possible to make positive statements about what it *is*?

Yes. But to do so, we must be modest. Much about how development works remains unknown, and we can't deduce it all—not even close—by a priori reasoning. If anything, the history of discoveries about biological development should tell us that. Instead, we should go much farther back to basics, to deduce some broad-swath generalities that—given the supreme caveat of the first law of adaptationism—generally or widely hold, or at least give us insight about what to look for in the design of developmental systems. The first law suggests that there are not likely to be any developmental properties or signatures that hold *invariantly* across all of biological development. However, the idea that there are tradeoffs that are likely to apply widely across developmental systems, coupled with the idea of form-function fit—that different systems will negotiate these tradeoffs differently based upon their history of selection and what they have been selected to do—can go a long way toward shaping the key elements in a conceptual framework for thinking about how evolved developmental systems work, and how to formulate and test evolutionarily meaningful hypotheses about them.

---

The first thing to note about the developmental systems that build organisms is that each has a design. I hope that by now I've done enough conceptual work that this term doesn't bother you, or at least you know what I mean by it. If not, all I mean is that developmental systems have properties, that these properties are what cause each system to generate the outcomes that it does, and that these properties exist for a reason: a history of selection acting on the variation that was present in populations through time.

The reason lies in Darwin's causal syllogism, the three elements that constitute natural selection: *If* organisms vary in their traits, and some variants are able to reproduce themselves better than others, and their design is re-instantiated, at least to a statistical degree, in their offspring—offspring resemble parents in their design—*then* natural selection occurs: Reproductively successful designs persist. This is a causal syllogism in that if each causal link in the chain is intact, then natural selection will and must occur (Endler, 1986). Developmental processes, in this syllogism, are the causal link between the design of a parent and the design of its offspring. They are what cause design to be re-instantiated, creating a positive regression between the phenotypes of parent and child.[2] What this means is that if natural selection occurs and is responsible for the design features of organisms, then there must be developmental systems that cause those design features to recur in each new generation. The logic behind this "must" is powerful and inescapable, and by itself is sufficient to show that the still-popular dichotomy between "evolved" and "developed" is not a dichotomy at all.

There are two entailments of Darwin's syllogism that are often forgotten or overlooked, but that are critical to a biologically sound model of cognitive development. One is that Darwin's syllogism entails that offspring resemble parents along dimensions that impact fitness, but it doesn't specify *how*. Classic models of evolution have tended to assume that all of the information needed to produce the resemblance between offspring and parents was carried by genes. We now know—and, I think, we could have known long before from the logic of the syllogism—that this isn't true (Jablonka & Lamb, 2005; Lewontin, 1974; Oyama et al., 2001; West-Eberhard, 2003). While genes are a necessary part of the causal chain that reliably produces design reincarnation across generations—they are a design feature of inheritance that was committed to early in the history of life (Maynard Smith & Szathmáry, 1995)—this neither means that they are sufficient to do so nor that they carry all of the information necessary to reincarnate resemblance. In fact, what produces the resemblance causally is the interaction between genes and everything else in the developmental process, both inside the organism and out (i.e., internal and external environments). This means, among other things, that whatever factors external to the genes that, in

[2] See Barrett (2006, 2007) and Tooby et al. (2003) for discussions of the notion of "design reincarnation"—a metaphor for the process of re-instantiation of design in each generation through developmental processes. The point of the metaphor is to focus on the primary function of developmental systems, which is to build phenotypes, and away from excessive focus on innateness per se. Crucially, developmental construction of phenotypes can involve processes and structures that are not present at birth, that change dynamically during the lifespan and vary across individuals, and that emerge through interactions between and within the developmental system and the environment.

interaction with them, reliably produce resemblance need not be coded into or informationally carried by the genes. The genes need only be selected to interact properly with that external stuff, whatever it might be—a form of inductive bet (Griffiths & Gray, 2001).

A second and related entailment of Darwin's syllogism has to do with variation. For natural selection to occur and for designs to persist over time, offspring must resemble their parents along dimensions that impact fitness, but they need not be identical. Darwin recognized this: Indeed, it is variation among offspring and between offspring and parents that provides grist for the mill of selection. In fact, if parents and offspring were exactly identical in every way, then selection could not occur. There is no reason why this couldn't be the case, of course, but while natural selection strongly favors replication fidelity in living systems, variation is hard if not impossible to eliminate. What this means is that developmental systems need not produce offspring that are identical replicas or phenotypic clones of parents, and under most circumstances never will. Instead, developmental systems will be selected for to the degree that they can reproduce the parental design along *fitness-relevant* dimensions.

This too has implications. It is true that the relationship between the designs of parents and the designs of offspring need only be a statistical one, and indeed, will almost *always* be only a statistical one, because all information-transmission systems, especially ones as astoundingly complex as the systems that transmit organismic design from generation to generation, have noise, or error. But this does not mean that just any statistical relationship will do. If we consider the phenotypic designs of parents and offspring in phenotype space and the statistical relationship that natural selection engineers as a mapping function between those shapes in phenotype space—which is what I am calling the developmental system—then there are certain generalizations about the kinds of statistical properties these mapping functions are likely, and unlikely, to have.

The most central property that natural selection will select for is replication fidelity along fitness-relevant dimensions. The reasons for this are stated perhaps most eloquently by Richard Dawkins in his book *The Selfish Gene* (1976) and supported by models of the "error threshold" in replication above which design breaks down over time (Eigen, 1971; Maynard Smith & Szathmáry, 1995). Dawkins envisioned a population of "replicators," which he thought of mostly as genes, but here we can think of the entire design of the organism as a replicator (indeed, this is how Darwin thought about it). Dawkins pointed out what at first seems almost paradoxical: that while evolution is change over time, and adaptations arise because of natural selection acting on variation, it is actually, at a certain level, variation that natural selection is acting to *prevent*. The phrase "at a certain level" is an important caveat, because natural

selection can select for systematic variation in designs, sometimes called adaptive plasticity (and it can even select for random development in some cases of unpredictable environments, known as "bet hedging"; Philippi & Seger, 1989). But, crucially, even in adaptively plastic designs, there is a level of invariance: namely, in the design features that produce the adaptive plasticity. If these are allowed to vary randomly, the ability to produce adaptive plasticity will degrade.

To better see Dawkins's point about why natural selection favors fidelity in design reincarnation, consider his idea of a replicator. Replicators are things that make copies of themselves. This ability is the property that makes both evolution and life more generally possible. Imagine a perfectly digital world in which replicators can produce exact clones of themselves and potentially create a population that is identical. But one kind makes exact copies of itself, and one kind systematically produces offspring that are slightly different. Which will do better? Or, in future states of the system, which type will be more common? It's easy to see that the type with self-replication fidelity will be most common, because the type that systematically produces variants will keep making offspring that are systematically different from itself in random directions in design space. Whatever design an organism might have, randomly varying it each generation as a matter of policy would be a guaranteed way to get rid of that design, as alluded to in the saying "if it ain't broke, don't fix it."

Now, you might say that much rides on my use of terms like "design" and "type." If we label the two types the "invariant" and "variant" types, then, assuming these reproduce at the same rate, won't there be equal numbers of the invariant and variant types at time t+1, t+2, t+3, etc? Yes. And in a world like the one Dawkins has set up, where there are no differences in individuals' ability to survive and reproduce, one could argue that the high-fidelity versus systematically variant designs have equal fitnesses when considered as abstract types. But our world is not a neutral, level playing field of that kind. Design features of organisms *do* impact their ability to survive and reproduce. And when you add this in, fidelity becomes crucial for preserving fitness-promoting aspects of design that natural selection has engineered. The reason, as both Dawkins and Darwin realized, is that not all variation is good for survival and reproduction. Indeed, most of it is bad.[3]

It is critically important to be clear, however, that this does not mean that natural selection favors crystallizing the design of organisms into innate, cloned, cookie-cutter copies that are identical in every detail. Above, I assumed a digital world in which exact phenotypic copying is possible, but even when we relax this

[3] Note that this is usually but not always true—there are some conditions under which high mutation rates can be favored, but these are special and probably relatively rare conditions (Sniegowski et al., 1997; Taddei et al., 1997).

assumption—as we must—Dawkins's point about replication fidelity remains, but in a slightly altered statistical form.

Remember that organisms are vastly complex, both genotypically and phenotypically. We can think of them as shapes in a space of a very high number of dimensions. When an organism reproduces, its offspring will not resemble it perfectly—sometimes due to noise or random error in the replication process and sometimes for reasons of design (when a woman gives birth to a male infant, for example, his phenotype is systematically different from hers, but by design). Crucially, not all dimensions and kinds of variation are the same, and we can make some generalizations.

The first, as we've seen, is that natural selection favors design replication fidelity along dimensions that are relevant to fitness. But design replication is not perfect; there is always error. Indeed, as Darwin recognized even before his theories were mathematically formalized, some degree of error, even a large degree, can be tolerated, as long as there is a reliable statistical signal, a correlation between parent design and offspring design that natural selection can latch onto. There are several ways of formalizing this statistical correlation and what it needs to be like for selection to work. One formalization is the concept of *heritability*, the proportion of phenotypic variation in a population that is due to additive genetic variation between individuals (Fisher, 1919).[4] If you think about the design of a parent as representing a point in design space, the offspring of that parent will appear as a cloud around that point—offspring need not be identical to the parent, but must be nearby in dimensions of design space where features are to be selected for. Otherwise, selection cannot occur.

A principle follows from this logic: While offspring designs can be a statistical cloud around the parent design, the cloud must be, on average, *centered* around the parent's design. Mutation is a force that tends to nudge fitness systematically downhill each generation, counteracted by the force of selection. But if parents produce offspring that systematically deviate from them in design space, and in a particular direction (i.e., not centered), then over time, the cloud will walk in that direction over the adaptive landscape. In general, the nature of

[4] What additive genetic variation means, intuitively, is variation that influences the phenotype independently of what combination of other genes the genes in question are inherited with. In other words, additive genetic variation reflects the average effect that variation in genes has on variation in phenotypes. In theory, this matters because genes are recombined each generation, so the net effect of selection on a gene will depend on its net effects on fitness when expressed across the variety of variations in which it can occur. Importantly, heritability is a concept related to genetic variation and defined contextually based on the amount of variability in the current population. It is not the same as the degree to which a trait can be inherited or passed on to offspring. For example, one can imagine zero genetic variation in a trait that is, nevertheless, reliably passed down to offspring.

adaptive landscapes is such that a directional walk of this kind might (if you're lucky) push you uphill for a little while, but it will also keep pushing you past it—unless selection kicks in. Thus, we do expect offspring designs to vary to a degree from the parental design, but only to the point where the (generally) large benefits of reducing such variation no longer outweigh the costs of doing so, which will determine the degree of scatter around the parental design. Crucially, these design variants will tend to occur in a statistical cloud centered on the parent's design with just this degree of scatter. Of course, organisms are objects of many, many dimensions in phenotype space, and this centering principle will only hold true along dimensions in which selection maintains design (along non-fitness-relevant dimensions, by definition, selection does not shape variation). We've so far only considered variation that's *random* with respect to fitness. What about variation that's not?

---

I claimed above that design reincarnation does not mean that natural selection favors offspring that are phenotypically identical to parents. I have mentioned at least two ways that this is the case. One is that natural selection will tolerate (fail to eliminate) some degree of scatter in design space around the parental design, as long as a design signal is there and the scatter is not systematically directional. Another is that many dimensions of variation have no impact on fitness, or not enough impact to counteract the other reasons for them occurring, such as what are sometimes known as developmental constraints, or unavoidable byproducts of the developmental process. But there is a third, extremely important reason that offspring may not be phenotypically identical to parents or to each other and yet that is not only tolerated, but actually favored, by selection: what is sometimes called "adaptive plasticity" (Stearns, 1989; Via et al., 1995; West-Eberhard, 2003).

You may have noticed that so far, I have switched back and forth between the concepts of *phenotype* space and *design* space. In order to understand adaptive plasticity, we need to make a precise distinction between these two concepts, because adaptive plasticity involves variation in phenotype space that is not the same as variation in design space. Rather, when we think about plasticity, phenotype stands in relation to design via a kind of mapping function: Phenotypic variation that is due to adaptive plasticity is the product of an underlying design to produce such variation (by "design," here, I mean developmental design, of course—it's also possible to think of individual phenotypes as designs). One way of thinking about such mapping functions is as the reaction norms that we discussed earlier in the book: different ways that organisms can develop depending on the circumstances in which

they find themselves (Schlichting & Pigliucci, 1998; Stearns, 1989; Waddington, 1957; West-Eberhard, 2003).

To understand plasticity and when it might or might not be selected for, let us first consider the question of what plasticity is. In the broadest terms, plasticity is the ability of an organism to change depending on circumstances. But remember that we are talking here about Darwin's syllogism and Darwinian dynamics: We are tracking phenotypes over time as they are transmitted from parent to offspring over generations. What plasticity means, in this context, is phenotypic variation or change over time, space, and circumstance (usually, within the lifetime of the organism). Most generally, plasticity will show up as phenotypic differences between a parent and its offspring, and/or between different offspring of the same parent; this is what plasticity means from the point of view of natural selection. To see that this is logically true, consider that if there are never phenotypic differences between parents and offspring—if they are always phenotypic clones—then there is no plasticity in the organism at all. Similarly, if there is no variation between parents and offspring along a particular phenotypic dimension, even if there is variation along other dimensions, then there is no plasticity *in that dimension*.

So far, so good. But as we've discussed, there are different possible reasons why we might see variation between parents and offspring along a particular dimension, and not all of these would be what we'd want to call plasticity—or at least, not a particularly interesting kind of plasticity. We defined plasticity as the ability to change in response to circumstances. Some kinds of changes, including many of the changes that would cause offspring to be distributed in a cloud around the parent's point in phenotype space, will be only very weak forms of plasticity. For example, if an offspring gets a limb chopped off during an accident in childhood, that is certainly a change caused by circumstances, but it is not the kind of plasticity that is of most interest to us, because it was random: It carries offspring away from the parental phenotype in phenotype space, but not in a systematically fitness-good direction. On the other hand, if the limb heals in response to the random cutting event, this is *not* an accident: It is an example of adaptive plasticity, plasticity that has been selected for *because* it puts the organism in a fitness-better part of phenotype space (healed wound) than where it was before (open wound).

What is the difference between adaptive and non-adaptive plasticity? At heart, the difference is the same as the difference between any adaptive versus non-adaptive aspects of organisms and follows the same underlying principle. To the extent that an instance of plasticity carries an organism systematically toward outcomes that are fitness-good, then that plasticity is likely to have been selected for, and therefore to be an adaptation. As elsewhere, when we are talking about fitness-promoting outcomes, it is of course possible that these can occur without any prior selection, and

it is indeed necessary that this happen, at least sometimes, in order for there to be anything to select for. But their tendency to happen systematically, in a given context, can ultimately only be preserved due to selection.[5]

This brings us back to why design space and phenotype space are not the same when it comes to plasticity. Imagine a developmental design that causes different phenotypes to be produced depending on circumstances. Fundamentally, the "scatter" of phenotypes that a given design produces when placed in different circumstances, its reaction norm, can be either adaptive (adaptive plasticity) or not (non-adaptive plasticity) (Stearns, 1989; Via et al., 1995; West-Eberhard, 2003).

Now, let's think about the directionality of the scatter—its statistical properties—and what selection will tend to do to it in both phenotype space and design space. Recall the arguments that I made for the statistical centering of design clouds. If designs pull systematically away from what's being selected for, selection tends to steer them back. This means that in *design* space, offspring design variations are possible but will tend to cluster in a centered cloud. In *phenotype* space, though, they might vary—and in the case of adaptively plastic reaction norms, they will, in systematic ways, even though they are products of the same underlying design.

To see that phenotype spaces and design spaces are not the same, consider an example. In humans, Chinese-speaking parents can produce children who speak English. These would count, on our definition, as different phenotypes. Imagine a case in which two Chinese-speaking parents who do not speak English produce a family of offspring all of whom speak English but not Chinese (i.e., the children are raised in a different language environment than the parents were). English-speaking and Chinese-speaking human phenotypes are likely to occupy at least some overlapping areas of phenotype possibility space: Being spoken languages, they have at least *something* in common, even if extremely minimal (though as we'll see, linguists have a devil of a time agreeing what those shared aspects, if any, might be; Evans & Levinson, 2009; Pinker & Jackendoff, 2005). But one thing about the relationship between Chinese-speaking and English-speaking phenotypes is certain: The language phenotypes of the English-speaking children are *not* a randomly scattered cloud centered around the Chinese-speaking phenotype of their parents. No matter how you map out English in phenotype space, it's not a cloud centered around Chinese; it's off somewhere else (some overlap, yes, but the less overlap you imagine,

[5] As usual, there are some exceptions to the idea that all forms of plasticity are selected for because they systematically increase fitness. For example, in the phenomenon known as evolutionary bet hedging, selection can sometimes favor phenotypes with lower-than-average fitness when there are tradeoffs between the mean and variance of fitness due to fluctuating environmental conditions (Philippi & Seger, 1989; Slatkin, 1974).

the more this makes my point). In other words, the children are all skewed in the same *particular* direction away from Chinese along dimensions of language phenotype space. And yet, the underlying design of the children's minds that allows them to acquire English instead of Chinese is not different in any systematic way from the design of the parents' minds that allowed them to acquire Chinese instead of English.

We will return to the example of language and language acquisition as a case study later. For now, the point is that reaction norms enact a kind of mapping function between developmental circumstances and the particular phenotypic outcomes they will produce in a given case. What plasticity means is that a given underlying developmental design can produce a sometimes startling array of developed phenotypes, depending on the circumstances. Crucially, the underlying design will often be relatively invariant in the sense I have outlined here: The design features that produce adaptive plasticity need to have replication fidelity in order to be selected for. But the phenotypic outcomes of this design might be wildly variant in phenotype space, depending on the mapping properties of the design. And no matter how wildly diverse the outcomes that mapping function produces might be, it will only produce systematically adaptive outcomes if it has been selected to do so under the various conditions it might encounter. This is a very important point to remember, and it will structure the rest of my discussion of developmental systems and how they work.

You might be wondering why I am going on about this distinction between variation in design space and in phenotype space at such length. It's because it is the crux of the problem with which I opened the book, and is at the heart of what muddles our intuitions about innateness and experience. How can it be, when I think movie thoughts or airplane thoughts or any evolutionarily novel thoughts at all, that evolution has anything to do with it? This is how: They can be what happens when you take an evolved reaction norm and put it in an environment with movies or airplanes. In fact, similar variation in phenotypic outcomes presumably existed even in the environments where human reaction norms evolved. Our ancestors living in southern Africa, for example, were thinking different thoughts with the same developmental machinery than human ancestors along the Mediterranean, or even than their neighbors next door (the details of where and when the relevant reaction norms evolved are, of course, of crucial importance, but they don't contradict the point that flexibility can be part of their evolved function). When we think and talk about invariants in development, then, it should be at the level of *design* space, the design of their mapping functions. And at the phenotypic level, we often, if not always, *expect* variation, even in features that are properly considered adaptations, for all the reasons outlined above.

I can't emphasize this enough, so I'll repeat it: At the phenotypic level, we often *expect* variation, even in features that are properly considered adaptations. What we should be talking about, in these cases, are the kinds of variation we expect, and why. Variation is not at odds with natural selection. In fact, when it occurs in systematically adaptive ways, it occurs *because* of natural selection. As we will see over the next few chapters, this is all too often misunderstood.

---

When can developmental systems be expected to produce adaptive phenotypic outcomes? Are there any general principles about the envelope of conditions under which adaptive reaction norms work, and/or what their mapping functions might be like?

Yes, there are. But, as everywhere else, the first law of adaptationism governs them. What we will see depends entirely on the details of the case, including in particular the available variation that natural selection had to work with—which depends on the evolutionary history of the structures involved and how they came to get there—and the costs, benefits, and tradeoffs that have determined how natural selection has acted in the given case. But also, as elsewhere, we can make some inferences about aspects of mind-world fit that we might expect to see, or not see, given particular kinds of tradeoffs.

The concept of an environment of evolutionary adaptedness is critical here. The EEA concept has been much maligned by critics of evolutionary psychology and even abused by some evolutionary psychologists themselves, particularly in cases where "the" EEA is treated as a single "thing" and associated with a particular time and place, such as the Pleistocene savanna (e.g., Kanazawa, 2004). Despite all the controversy and abuse, however, an evolutionarily correct version of the concept is both logically unavoidable—every adaptation has an EEA—and indispensable for understanding what adaptations are adaptations *to*.

So what is the evolutionarily correct version? You might have noticed that all the considerations I gave above about how design evolves and reproduces itself in design space and phenotype space have to do with evolutionary history: the historical trajectory that populations of phenotypes have taken across fitness landscapes to end up where they are today. As evolutionary biologists emphasize, these trajectories sum up an enormous multitude of causes and effects acting over space and time—including but not limited to natural selection—to produce the phenotypes that we see.[6] The

[6] Technically, the net fitness of phenotypes experiencing environments that vary over time is not a straight sum of individual fitnesses, but rather is proportional to the geometric mean of those fitness effects across the environments they experience (Orr, 2009). The reasons are complicated, having to do with nonlinearities in the effects of current events on future generations. The take-home point is that fitness is not always "averaged" across environments in the strict mathematical sense; the number-crunching can be more complex.

pathway that a population took over the fitness landscape during the course of its evolutionary history tells us *why* members of that population now have the properties they do: morphological, behavioral, developmental, genetic, and all other properties as well. And at every point in this evolutionary process, the pushings and pullings of replicating phenotype clouds across the landscape had to do with interactions between the current phenotype and the current environment. Take away an understanding of these interactions, a correct description of them, and you take away an understanding of why the outcome of history—what we see now—is what it is.

This is where a proper understanding of the EEA concept lies: in the *interaction* between phenotypes and environments, aggregated across space and evolutionary time. The EEA of any evolved feature of an organism is, quite simply, the set of environmental properties that interacted with phenotypic properties of the population over time to shape the evolution of that feature (Tooby & Cosmides, 1992). Note, importantly, that different features, or adaptations, have different EEAs. An EEA is neither a place nor time, but rather a long smear of events occurring over space and time. If we could see in four dimensions, perhaps we could visualize it as a long, fluffy, worm-like cloud.

What sometimes disturbs people about the EEA concept is that it refers to a *subset* of environments or environmental properties, rather than all of them. This sometimes seems to have an air of subjectivity about it: either subjectivity on the part of researchers or theorists, or illicit personification of natural selection, as in the metaphor that natural selection only "sees" certain parts of the environment. However, it's a mistake to reject the logical necessity of EEAs because of the possible sources of human error in understanding or interpreting them. To be sure, there is an epistemological problem in being sure that we've identified the correct set of environmental properties that played a causal role in the evolution of any given feature of an organism. But such epistemological problems bedevil *any* account of causation, and so are hardly unique to the EEA concept.

It is a logical necessity that every evolved feature of every organism has an EEA, just as it is a necessity that the EEA of any given trait is a subset of all possible environmental properties. This follows from the fact that only some environmental properties, in some places and some times, entered into the interactions that shaped the evolution of any given feature of an organism. And this, in turn, entails that the EEA concept is a historical one, having to do with the causes that led to a particular organismal feature being the way it is now. And even if an adaptation is not currently in an evolutionary equilibrium or stable state (which could be true of many, if not most, of an organism's features), it is the case that there is some set of past and ongoing selection pressures that has given rise to its current design.

I will put aside the epistemological problems of knowing whether we've gotten the EEA right in a given case for another time, because here I am focusing on the logic of mind-world fit and what principles we can derive about it. What does the fact that developmental systems were shaped by a history of interaction with particular environments over space and time tell us about the possible designs of those systems? And what does it tell us about the question posed above, namely, when can developmental systems be expected to produce adaptive phenotypic outcomes?

The most basic principle follows from the fact that developmental systems, like any adaptations, have been shaped by a process that whittled down the available pool of variants into those that worked well in a specific (statistically distributed) set of environments. Therefore, these systems can be expected to produce the most adaptive (fitness-good) outcomes when the environments they currently occupy match their EEA along relevant dimensions. It is hard to imagine a simpler or more logically obvious principle. The logic of the principle follows from the fact that the trial-and-error process of natural selection happened in some set of environments that actually *occurred*, and no other. Therefore, although it is possible that the designs produced by this process *might* produce outcomes that are even fitness-better in certain environments that the population hasn't yet experienced, the only ones that are guaranteed to reliably or systematically produce fitness-good outcomes are ones that resemble those in which the design features of the adaptations were forged. Outside of that statistical envelope of conditions, all bets are off.

This does not mean that we couldn't find some previously unexperienced environments where an evolved design would do better than in the conditions under which it evolved. Indeed, we could almost be guaranteed to find some fitness-better ones, if we knew enough about the design in question and how it worked. What we mean by the "statistical envelope" of conditions that have shaped an adaptation is a difficult and complex question that we'll be revisiting and refining in the coming chapters. Importantly, there will, by definition, have been no selection for mechanisms to cause an evolved design to find fitness-good areas of environment space that are entirely novel. Why "by definition?" Because if there has been such selection, then similar conditions must have existed in the design's EEA and are therefore not novel. If the properties of novel environments are such that mechanisms can exploit them in systematically good ways, it's because the inductive bets that held in ancestral environments still hold in the new ones.

So we have a general principle: Adaptive outcomes are most likely to occur when a developmental system operates in environments that are similar to its EEA along

relevant dimensions.[7] Of course, you might have noticed that certain terms need to be further defined—in particular, terms like "similar" and "relevant." And it is in the unpacking of these terms that the most interesting design considerations and tradeoffs emerge.

---

Consider the possible diversity of pathways between what an organism inherits—its DNA, epigenetic modifications to its DNA, its internal biomolecular environment, its social inheritance, its physical environment—and how the organism develops phenotypically. The potential array of such mapping functions, or reaction norms, is staggering, virtually impossible to imagine. Over the evolutionary history of any lineage of evolving organisms, out of the giant set of *possible* mapping functions, some subset—tiny in relation to the whole possible set, yet still enormous—has actually appeared, and this is the subset that natural selection has acted on. Out of all the developmental mapping functions that have appeared, natural selection has favored the ones that worked fitness-best in the environments that the lineage experienced.

Are there generalizations about how the shapes of these mapping functions relate to the shapes of the world in which they evolved, and to the shapes of the fitness landscapes that whittled them down over evolutionary time? When it comes to development, this is the central question of mind-world fit. Are there general principles of goodness of fit?

In principle, natural selection can generate just about any developmental mapping function. It can create reaction norms that are entirely flat, where the same phenotypic shape is produced across the entire range of circumstances an organism might experience, or reaction norms in which phenotypic outcomes vary contingently with developmental circumstances down to almost any degree of precision. It can in principle create any shape of reaction norm: straight lines, step functions, linear functions, or curves of great complexity. Any of these is possible theoretically. What we actually see, what natural selection actually produces, depends on the variation in mapping functions that has been available, and the various costs, benefits, and constraints that have determined the shape of the fitness landscape on which that cloud of variation has been pushed around. Again, we see the first law.

But this doesn't mean we should throw up our hands and give up—far from it. It means that we might be limited in our ability to make predictions in all possible

[7] See Irons' related concept of Adaptively Relevant Environments, or AREs (Irons, 1998).

cases, but it doesn't mean that we can't understand the design considerations that might apply in a given case, explain designs we see, or even make guesses about designs we are likely to see. Let us consider a few hypothetical shapes of developmental mapping functions, and how and why they might evolve.

One kind, as I mentioned, is a flat reaction norm, which produces the same phenotype regardless of circumstances. This is sometimes called a canalized design (Waddington, 1957). Canalization doesn't mean that the phenotypic outcomes produced by such a design need to be identical along every dimension, just along the dimensions that are canalized, where the reaction norm is flat. Nor does it mean that even along canalized dimensions there won't be *some* phenotypic variation (e.g., unavoidable developmental noise). It just means that there is no systematic, designed response of the phenotypic outcome to varying circumstances.[8] Instead, flat reaction norms reflect either that a single kind of phenotypic outcome has been favored across the range of developmental conditions, or that no variants that were sensitive to that dimension ever arose for selection to act on.

When might flat reaction norms evolve? Most obviously, they might occur when the EEA is flat along a particular dimension—there was no variation in that dimension of the environment across ancestral time and space—*and* when the fitness landscape was such that a particular design was either favored for that setting of the environmental dimension, *or* when constraints of evolutionary history restricted the population variation to some portion of the landscape, or a combination of both. It's important to note here that most environments probably vary in most dimensions you can think of, so in the space of all possible dimensions of variance and constancy, the constant dimensions are likely to be a special subset. Nevertheless, there certainly are some dimensions of constancy or near-constancy in the environments of most species. Obvious examples in our case include things like the laws of physics and physical constants, universals like the existence of sunlight and objects, the difference between liquids and solids, and so on. And there are likely to be many other higher-order invariants as well. For example, every environment contains predators, but the predators in any particular environment vary. Every human environment contains spoken language, though the details of the languages differ. Every human

[8] Note that there are important ways in which one might consider staying the same despite varying circumstances to be a "response." Selection is usually required to maintain flat reaction norms, and specialized mechanisms can be necessary to ensure development turns out the same across many conditions, by steering it toward that outcome from many different directions. By analogy, lots of money and specialized design features are necessary to keep skyscrapers and bridges from "responding" to high winds and earthquakes. Interestingly, these can include forms of homeostasis-like plasticity to keep the important design features intact.

environment contains individuals whose behavior is caused by mental states, but the contents of those mental states are enormously variable. And so on. Consistent with this, we and many other species are likely to have many, many design features that exist because of these universals, ranging from non-psychological traits like the way that our bones and body morphology are designed to deal with constants of gravity, to psychological ones like the use of sunlight and assumptions about its properties by the visual system, to assumptions built into learning systems about what is likely to be present in their learning environments, which we'll see below. And in addition to fitness optima for dealing with these constants, there are also constraints of evolutionary history: Both human lens eyes and fly compound eyes rely on constancies of light, but do so in different ways, presumably because of the historical pathways that led each lineage to separate solutions to the adaptive problem of using light to gather information about distal objects.

In general, species-typical phenotypic universals are likely to be most often explained by some combination of the considerations given above: constancy along some dimension of the EEA, plus a peak in a fitness landscape *or* historical constraint to some region of the landscape or, almost always, both. But remember that when we talk about flat reaction norms, we are only talking about *phenotypic* universals. There are many species-typical *design* universals that result in phenotypic variation that we have yet to consider. How much ink has been spilled about "human universals" from the failure to understand this difference between phenotype space and design space?

---

So far, I've talked about flat reaction norms evolving in flat regions of environment space, where there is no environmental variation along a dimension that could select for a plastic developmental response. But it is also possible to get a flat reaction norm even when there *is* environmental variation along a particular dimension. There are several situations in which this might occur.

Remember that we are talking about (at least) three kinds of spaces. There is environment space and phenotype space. The reaction norm is the function that maps between them for a given developmental design (in most accounts, this is synonymous with genotype, but remember that we must also allow for inheritable non-gene elements, such as epigenetic effects). Then there is the fitness space, the shape of which is the fitness landscape. A given reaction norm in a given set of environments will lead to a set of phenotypes, and these will have fitnesses, the distribution of which determines how selection pushes the shape of the reaction norm around in design space. Given these considerations, how might it be possible to get flat reaction norms in variable environments?

One way, the most obvious way, is that, in the region of environmental variation in question, the fitness landscape has a peak or plateau for a single phenotype, even across differing environments. Another way of saying this is that it's best to be a single way, the same way, across a range of environmental variation. There are plenty of cases in which one might imagine this could be true. To give a metaphoric example from human-made design, wheels might be a good solution for a variety of terrains. For a vehicle that has to drive on asphalt, gravel, and dirt, it could be that the wheel design (as opposed to, say, legs or tractor treads or skis) is the best design, separately, in each environment. In that case, selection will favor a single phenotypic outcome, a flat reaction norm, across the environments of asphalt, gravel, and dirt. This is sometimes called "counter-gradient selection" or "antagonistic selection" (Conover & Schultz, 1995; Falconer, 1990; Nylin & Gotthard, 1998). For example, selection seems to have canalized growth rates of some animals that live in environments with variable growing seasons, speeding up growth in short seasons and slowing it down in long seasons so a similar adult size is reached regardless of environment (Conover & Schultz, 1995).

We might see examples of such situations in how natural selection chooses among computational designs as well. For example, a Bayesian design might be optimal for updating probability estimates across many circumstances. If so, a Bayesian design might have been selected for many cognitive systems, across a range of developmental environments, and one might expect to see a flat reaction norm in the developmental systems that build Bayesian properties into developing cognitive phenotypes.

However, while flat reaction norms can be selected for when the same design is "best" across the entire range of variation the design encounters in its EEA, there are conditions under which flat reaction norms can be selected for even when the same design *isn't* fitness-best across all the environmental conditions it encounters. How can this be? In essence, selection can "settle" for a less-than-optimal phenotype in some regions of environment space when the costs of producing that phenotype developmentally—the costs of having a suitably plastic reaction norm—outweigh the benefits of the added flexibility. This might occur, for example, when a particular phenotype is the best one across the majority of the species' environmental circumstances. In these cases, selection can "split the difference," picking a flat reaction norm that produces the single phenotype that does best, on average, across the range of the species' environments. In fact, it's theoretically possible (and perhaps even common) to have a suboptimal phenotype across the entire environmental range if selection against a plastic norm is so strong and the fitness landscape is so rugged that very few organisms find themselves exactly on peaks, although they tend to be closer to the peaks than the troughs.

Finally, keep in mind that while I have been talking about flat reaction norms that evolved due to selection, there will also be many reaction norms that are flat just because no variant that plastically responded to a particular environmental dimension ever appeared in the species' history. Bacteria, for example, don't respond to readings of Shakespeare; their reaction norm in the literary environment is entirely flat. This is "constraint" in the sense that no ability to have a response ever appeared. Humans undoubtedly have many flat reaction norms (failures to respond to conditions that in fact vary) for reasons of this kind.

How common are flat reaction norms? This is an interesting question to which I don't think we know the answer, but there are some things we can say about the reasonableness of even asking the question. Organisms are massively multidimensional shapes in phenotype space and developmental space. We haven't sampled nearly enough of the human phenotypic shape across human environments to be able to assign anything like a number to the relative "prevalences" of flat versus non-flat reaction norms. But here is something that is definitely true: There are lots of ways to carve up or measure a massively multidimensional shape. A lot depends on how you lay the measuring tape across it. Biologists, being practical, are in the habit of picking particular carvings or dimensions of measurement and calling them "traits" (or "features" or "characters")—which is fine as long as one doesn't forget that any such carving is necessarily arbitrary.[9] For any aspect of an organism you look at, the amount and kind of variation you get can and often will depend on the dimension(s) along which you measure. This is a kind of frame problem, but not for selection—for us. You could call "fingers" a trait, but for human fingers, there are limitless numbers of dimensions of variation (how long they are, how wide they are, their color, number of hair follicles, etc.), as well as limitless numbers of dimensions of invariance (how many of them humans typically grow, which end the fingernail is on, the fact that they have bones inside, the chemical building blocks of bone, highly conserved regulatory systems for building limbs, etc.). In discussions about human "universals," this point is often missed: One side points to variation and the other to constancy, without appreciating that whether you see variation or constancy depends at least in part on how and what you measure.

Perhaps not coincidentally, flat reaction norms seem to be at least one part of an intuitive notion of "innateness": Sometimes when people say that a trait is innate, they mean that it must be identical in everyone (Ariew, 1999). Certainly this is one

[9] By "arbitrary," I mean that most empirical measurements, which are products of particular measurement technologies, have a degree of arbitrariness to them. It is in principle possible to have a non-arbitrary definition of a trait: for example, by defining traits in terms of modularity (Wagner & Altenberg, 1996; Wagner et al., 2007).

way you could think about innateness if you want. But I do hope you can see that innateness in this sense is *not* the same thing as phenotypic outcomes being "caused by genes" or "caused by evolution," since the outcomes of plastic reaction norms are *also* caused by genes and by evolution, just like flat ones. Some people might like to solve this problem by just thinking of the reaction norm as innate, whether or not it's flat, which is a defensible position. But if you do that, you have to agree not to point at some *phenotypic* outcomes as innate and others not (a temptation that might be difficult to resist).

For the sake of argument, let's imagine that flat reaction norms shaped by selection are actually rarer than non-flat ones. Because of the problem of measurement frames, I'm skeptical that one can make much sense of such a claim, but let's say we figured out a way to representatively sample reaction norms across the many-dimensional space of an organism. If we could do so, it's plausible, especially for complex organisms like humans, that entirely flat reaction norms are rare—meaning, it's probably relatively difficult to find traits that come out *exactly* the same regardless of developmental environment. If so, and if you insist on equating the term "innate" with flat reaction norms (i.e., canalization), then you are forced to conclude that most phenotypic traits operated on by selection are not innate. This is a conclusion, I suspect, that even those who insist on using "innate" that way would not like to reach. In my opinion, this is a fairly good reason to stop using "innate" to mean "evolved" (or more precisely, to stop equating "innate outcomes" with "evolved outcomes"). It's also a good reason to stop thinking and talking about innateness and to start thinking and talking about developmental design, and how natural selection has shaped developmental systems to produce phenotypes given environmental circumstances.

---

While there is only one way to be flat, there are lots of ways to be non-flat. The set of possible mapping functions from environments to phenotypic outcomes is enormous. But remember that when we talk about plastic reaction norms, we're necessarily talking about evolved developmental *responses* to environmental variation. If it's not a response—if there is not a systematic mapping function between environments and phenotypes—then we are talking about random outcomes, which are outside the scope of natural selection and therefore of design. Moreover, when we are talking about *adaptive* developmental responses to environmental variation, then there must have been a history of selection to produce them; for reasons outlined above, adaptive responses only occur systematically within the envelope of EEA-like conditions. I stress this not just so that we keep in mind the relevant forces shaping mind-world

fit, but also to steer away from the common tendency to think that when we're talking about plasticity, we're not talking about adaptations.

As I mentioned, the potential variety of shapes of possible mapping functions from environment space to phenotype space is staggering, and which ones actually evolve depends on the baroque details of costs, benefits, and landscapes: the first law. Given this, there are likely to be fewer generalizations one can make about the whole enormous class of non-flat reaction norms than one can make about flat ones. But I will attempt a few.

Perhaps the simplest kind of developmental plasticity involves what biologists call *morphs*—discrete types or phenotypic designs within a population. These, in turn, come in at least two kinds: developmentally fixed morphs and developmentally facultative morphs (West-Eberhard, 2003). Developmentally fixed morphs are usually thought of as "genetic" morphs: types within a population that are discrete types because they carry different DNA (also called polymorphisms; Stearns, 1989). In humans and many other sexually reproducing species, the most obvious morphs of this kind are males and females: Humans born with two X chromosomes develop into the female phenotype, and humans born with an X and a Y develop into the male phenotype. I say "develop into" males or females to emphasize that human males and females are, in terms of the genes and developmental systems they inherit, virtually identical, or at least very highly overlapping. Both sexes have almost all the DNA required to develop into either males or females, the critical difference being the possession of DNA on the Y chromosome that will make you male, if you have it. This is oversimplifying a bit (Sinclair, 1998). But the basic point stands: There is a kind of switch, determined by a little bit of DNA that you are born with, that turns you into one of the two possible morphs. This falsifies a common misunderstanding about the relationship between genes and phenotypes, namely, that large differences in phenotypic design between individuals require corresponding large differences in DNA *between* those individuals. This isn't to say that the genes that make us male or female are all the *same* genes; many different genes are involved in creating the differences, even though we all carry them. But differences between the sexes—even large differences in phenotypic design—can be and largely are the result of how the same genes are *regulated*, turned on and off and interacting during development, rearranging how particular developmental resources are used. We will return to this issue later when we discuss the question of how much genetic difference between species is required to produce design differences.

When you define morphs in terms of discrete types, males and females are one of the only major developmentally fixed morphs we know of in humans—though elsewhere in the biological world, there are plenty of other kinds, including, for example, insect morphs such as workers, soldiers, queens, and so on. But when you relax the

requirement of phenotypic discreteness and allow continua, many kinds of human variation caused by differences in genes are continually varying morphs of a sort. One kind, for example, might be the continuum of personality, with evidence suggesting the existence of different personality dimensions and factors like one's propensity to take risks at least in some part determined by differences in DNA (Chen et al., 1999; Jang et al., 1996). Others might be continua in geographically distributed phenotypic traits affected by genes, including traits like skin color or ability to digest certain compounds (Novembre & Di Rienzo, 2009; Perry et al., 2007; Sabeti et al., 2007). Of course, genetic differences *between* individuals aren't examples of developmental plasticity at all: They are caused by inheriting different reaction norms, not the same one that produces different outcomes depending on environment. This is, in fact, what it usually means to say the differences are genetic.

What about facultative morphs? These are sometimes called *polyphenisms* to highlight the fact that discrete phenotypes are being produced by the same underlying genotype (Stearns, 1989). You might have thought that sexes are necessarily genetically fixed morphs determined by the combination of sex chromosomes you inherit at birth, but this isn't necessarily so. In some species of animals, including certain coral reef fishes such as the bluehead wrasse, the same individual can develop into a male or a female depending on the circumstances, and can even *change* from one sex to the other depending on rapidly changing factors such as the mix of other sexes in the local population, a phenomenon known as sequential hermaphroditism (Warner, 1984). In humans, this doesn't happen in the case of sex, though it is possible to think of certain developmental change, such as the change from pre-reproductive juvenile to reproductive adult, as being a change from one morph to another, with evolved developmental systems causing change in both morphology and behavior (driven by changes in gene expression during puberty). In fact, there is some evidence that the timing of this transition from juvenile to adult in humans—the onset of puberty in girls—is affected by aspects of the local social environment, in particular, the absence or presence of parents and their positive or negative behavior toward the child (Belsky et al., 2007). So the concept of discrete developmental changes being triggered by changes in the social environment is perhaps not so unusual after all.

Developmental morphs are sometimes thought of as developing via a kind of switch: The reaction norm is such that, depending on circumstances, one morph or another is produced (Stearns, 1989). This differs from what I called "fixed" morphs, or genetic morphs, because in developmental morphs, the same individual has the capacity to reach multiple phenotypic points—morphs—and which one it reaches is determined by the developmental mapping function that maps between circumstances and morphs. There are many potential kinds of such mapping functions that vary in the shape of the curves that map from circumstances to morphs. Some, for

example, might require a buildup of experience over time in order to commit to one morph or another, whereas others might be activated by an immediate trigger. In the case of the bluehead wrasse, for example, upon death of the largest male in the group, the largest female will turn into a male within a period of several days (Warner, 1984). Which developmental design evolves depends on the cost/benefit tradeoffs of committing to one morph or another, given the certainties or uncertainties that being that particular morph will pay off, and the costs of retooling from one morph to another (Frankenhuis & Panchanathan, 2011; Munday et al., 2006).

One might think that "triggered" developmental morphs of this kind are rare or nonexistent in humans, and indeed that is the standard view. But the example of puberty, when self-generated changes in the internal developmental environment (as well as some external events, such as parent absence) cause developmental processes to turn on and off leading to changes in morphology and behavior, hints that the answer might depend in part on how we choose to define things like "triggers" and "switches." In fact, genes turning on and off is not only common in humans and other animals but the norm, at a variety of timescales from the developmental (e.g., turning genes on and off to produce tissues of different types) down to the immediate short-term timescale of how cells go about their daily business (e.g., turning genes on and off to digest specific nutrients; Jaenisch & Bird, 2003; O'Brien & Granner, 1996). In the bacterium *Escherichia coli*, for example, the *lac* genes produce an enzyme and a transport protein involved in the digestion of the sugar lactose; these genes are activated by a system of switches that turns them on when lactose is present in the environment and turn them off at the end of the meal, changing their phenotype from a digestive to a non-digestive one (Jacob & Monod, 1961). In humans, the protein insulin, which is produced by the pancreas when proteins and glucose from digested food are present in the bloodstream, regulates the expression of over 50 other genes (O'Brien & Granner, 1996).

What about brains? While methods are not yet available to study moment-to-moment gene expression in working human brains, we do know that thousands of genes differ in how they are expressed across the brains of mice (Lein et al., 2007) and that specific genes are turned on in mouse brains during and after learning (Guzowski et al., 2001; Hall et al., 2000). In honeybees, expression of different sets of genes in the brain is linked to specific types of behaviors, like nursing and foraging (Ben-Shahar et al., 2002; Toth & Robinson, 2007; Whitfield et al., 2003). Given the expansive definition of phenotype that I've suggested we adopt, there may be many changes in human psychological phenotypes, even at short timescales, that are the result of switches being turned on or off. Mood might be an example, but so might be things like learning. In psychology, it's standard to distinguish between "developmental" and "psychological" processes based on timescale—developmental processes are

longer-term—but when thinking about plasticity and the developmental systems that regulate the phenotype based on current circumstances, there may not be a natural dividing line between developmental and psychological plasticity. The regulatory systems that turn things in your brain on and off depending on circumstance—for example, in happy times or sad times, hungry times or full times—may be examples of plastic developmental systems that operate over relatively short timescales. There is no question that this is a nonstandard use of terminology, but it might make sense to think of ourselves as containing a variety of morphs—what might be called *contextual morphs*—that we can slide into or out of, like Jekyll and Hyde, as circumstances dictate. Such phenotypic states would, of course, have reaction norms, and would have to obey the rules of mind-world fit if they were to have evolved.

---

So much for phenotypic discreteness. What about more continuous outcomes? So far, I've focused on relatively phenotypically discrete outcomes where selection has created a mapping function between particular circumstances and particular outcomes. If you imagine the shape of those functions, you've probably been thinking of something like a step function or an if-then statement: In situation A, become X, and in situation B, become Y. But as I mentioned, there can also be a whole variety of curves of any imaginable shape. For example, there is a reaction norm that maps between food intake over the course of your childhood as a continuous variable and your adult height as a continuous variable. The relationship is a curve (Bogin, 1999). You can also imagine, perhaps more abstractly, a mapping function between some aspects of childhood environment and adult disposition, or personality, or indeed any trait.[10] Hopefully, by now, you have no problem imagining such mapping functions and envisioning how evolutionary processes might shape them.

But many psychologists might say that these are all examples of relatively "closed" developmental programs, where, although there is plasticity, the mapping between situations and outcomes is already set. For every environmental situation, the outcome is known. What about more "open" programs, where the outcome hasn't been *specifically* selected for in advance? For example, children raised in an environment

[10] Biologists Carl Schlichting and Massimo Pigliucci (1995), for example, distinguish between "phenotypic modulation," in which phenotypic outcomes are a smooth, continuous response to some environmental variable, and "developmental conversion," or control of development via gene regulation, which can lead to very complex mappings between environments and outcomes. The results of developmental conversion needn't be smooth and proportional responses to environmental variation but can involve step functions (switches) and even more complex and nonlinear environment-phenotype mappings.

where they read the novel *Moby Dick* grow up with a phenotype containing detailed information about Ahab and Queequeg. Surely there is not a pre-specified reaction norm that maps between *Moby Dick* environments and Ahab/Queequeg developmental outcomes?

This is a complicated question. I will suggest that the answer is ultimately yes, because the word "pre-specified" is a red herring. But it will take some time to unpack this—indeed, the whole next chapter is devoted to this problem. Before we get there, let me sum up a few things about the developmental view I've been building.

First and foremost, I hope I've convinced you that it's important to think about development in adaptationist terms: Developmental systems are things that natural selection builds because of the phenotypic outcomes those systems produce. This means that when we think about developmental systems, all the rules for thinking about adaptations apply. For example, they have proper and actual domains. Each developmental system is adapted to a particular set of conditions, its EEA, and will only tend to produce systematically adaptive outcomes when its current environments are matched to the EEA along the dimensions it is designed to handle. We are also forced to consider what the design features of developmental systems are: what properties they have that cause them to produce systematically adaptive outcomes when placed in the right conditions. As elsewhere, there is an enormous scope for what those design features might be, and the first law applies. But there are no developmental systems that produce adaptive outcomes "just because." They only do so in virtue of adapted design.

Second, I hope I've convinced you that the distinction between what's innate and what develops is not a particularly cogent or useful one. It's common to think that babies are born with a certain endowment that is innate and shaped by evolution, and that all changes after that are due to development, which is shaped by experience. What I hope I've convinced you of is that while there *is* a phenotype that is present at birth, and while there *is* learning, the idea that the menu of possible developmental designs consists of just these two ingredients, like selecting a meat and then the sauce to put on it, is an impoverished and biologically unrealistic one. Birth is just one point in the developmental process, and genes must play a role in the development and operation of the brain throughout the lifespan. Why not consider the whole developmental process as something that natural selection shapes?

A third and related point has to do with rethinking the role of genes in development. It's not that genes aren't critically important—they are. But the ideas that many of us have about the role of genes, what they actually do, are likely to be wrong. As far as we know in our current state of understanding, the genetic system is best thought of as a very complex computational machine that guides the process of development. Genes don't simply provide a picture of phenotypes to be

built, nor do they do *anything* in the absence of interaction with environments, internal and external. As we will see later, this renders certain common ideas about the relationship between genes and mental structures—such as that the psychological differences between any two species should be linearly proportional to the genetic differences—flat wrong. For example, it's now abundantly clear that the vast majority of genes you need to build a human are also in a chimpanzee, and for that matter, a mouse. Of course, there are genetic differences, and some combination of differences in genes and environments, including social environments, must account for human-chimp phenotypic differences. But a large part of the story must involve differentially turning genes on and off at different times and places in developing tissue (Carroll, 2008). We don't yet have a handle on how rearranging a very similar toolkit leads to the differences we see, but we're beginning to find out (Blekhman et al., 2008; Caceres et al., 2003; Enard et al., 2002; Khaitovich et al., 2004, 2006).

Finally, the view I'm advocating is inherently an additivist one. Although we can and often must think about things like reaction norms as "chunks," organisms are in fact wholes, and it is crucial not to lose sight of the view of both phenotypes and developmental systems as massively multidimensional shapes, where all parts of the phenotype and all parts of the developmental system are in fact parts or regions of the same object. What progressive changes to developmental systems do is to change the shape of the developmental object so that new outcomes are produced. These outcomes are always changes to older systems—nothing is produced entirely anew in this sense—but it is also the case that when these changes are selected for, they can add something new to the developmental repertoire. Even in cases where we can view evolutionary changes as "taking something away" (e.g., making a previously complex reaction norm less complex), they are still, if favored by selection, allowing the organism to reach areas of fitness space they could not reach before. More often than not, however, this will be because the changes allow the developmental system to produce new phenotypes that it could not produce before: phenotypes that can, for example, respond to aspects of the environment to which they were not previously responsive. When it comes to mental development, what is most interesting—and what, I will argue, requires some new ways of thinking about what developmental systems do—is that these changes are not achieved by adding entirely new organs, like fingers, but by rearranging the structure of an existing organ: the brain.

# 7

## OPEN ENDS

Just now, I was thinking about how slow the computer on which I'm writing this is, and how I need to replace it with a new one soon. When it comes to the kinds of thoughts that pass through my head on any given day, there doesn't seem to be anything particularly special about this one or my ability to entertain it. Humans are able to think about computers, reality shows, *Moby Dick*, kayaks, differential equations, and pasta. They are able to look up the time that a particular movie is showing and plan their day around that. They are capable of designing and building the computers on which they look up those movie times. They are able to think and argue about whether *Taxi Driver* or *Apocalypse Now* is the better film. And they are able to think and argue about where the abilities to do these things came from. There seems to be no limit to the variety of thoughts humans can have, and the scope of what we can think seems to far exceed that of any other animal—at least as far as we know.

I claimed at the beginning of this book that evolution is responsible for this. I claimed that evolved mechanisms enable and frame our every thought, including those mentioned above. I've also said that developmental systems have been selected for *because* of the outcomes they produce, or more specifically, because of the outcomes they produced in the past.

But these seem incompatible. Am I saying there are developmental systems that have been selected to produce movie-thinking phenotypes? How is that possible if movies didn't exist in the EEA? And given that reaction norms map between developmental circumstances—environments, experiences—and outcomes, how could any reaction norm have "thinking about whether *Taxi Driver* or *Apocalypse Now* is the better film" as an endpoint that was ever selected for?

I won't, you might be pleased to hear, be suggesting that any such outcome has ever been specifically selected for. But I will suggest that deciding which of two movies is better is a *token* of a *type* of outcome that was selected for, and therefore, that the developmental systems responsible for producing such a thought are functioning just as they were selected to do. In order to see how (or if) such a claim makes sense, however, we'll need to rethink exactly what we mean by a reaction norm as a "mapping function."

A typical way to think about "functions" is the way we learned in math class: a function, as in f(x) = y, maps from one variable (x, in this case) onto another variable (y), where each value of x maps onto one value of y. In our calculus or algebra class, these variables are typically thought of as values along some linear dimension, as when x and y are real numbers. A linear axis is all that is required to define each and captures everything we need to know about those variables; specify x, and you get y.

In the last chapter, I talked about reaction norms in this way and stressed that just about any shape of mapping curve was possible, from perfectly flat to fantastically complex, and that natural selection adding up the costs and benefits of different variants of curves over time would produce the reaction norms we actually see. In this formulation, the input variables, x, are developmental circumstances (environments and experiences in those environments) and the output variables, y, are phenotypic outcomes. For the development of many traits, it's easy to envision such mapping curves: For example, a curve that maps between the amount and quality of food you ate as a child and your eventual adult height, or between the amount of interaction with animals you had as a child and your adult fear (or lack thereof) of dogs. But what kind of mapping function has as its endpoint the ability to think about the mental states of others? Or, for that matter, what kind of mapping function has as its endpoint *any* mental capacities with variable content? Are there really reaction norms whose endpoints include playing chess, driving, writing books, and arguing about movies? And if so, how could we possibly talk about those reaction norms as mapping functions and document their properties?

For linear or dimensional traits, where we can map them along a scale or continuum, it's possible to imagine pre-specified relationships between environments (e.g., food intake) and outcomes (e.g., height).[1] This is because for a variable that has just a single dimension, like height, there are no additional dimensions or parameters: Your height is just a single value, period. In these cases, it's easy to imagine a mapping function with all the values, all the input-output relationships, specified, as in the logistic growth function $y = 1/(1+e^{-x})$. Here I'll call this kind of curve-like mapping function between one unidimensional variable and another a "classical" reaction norm, because such norms (mapping between one environmental variable and one phenotypic variable, mediated by genotype) have been most studied in biology

[1] I am allowing for the sake of argument that it makes sense to talk about "pre-specification" in this case, but I'm not convinced that's true here or anywhere else in biology for that matter. Pre-specification, like innateness, only makes sense in terms of inductive bets, because developmental systems will only produce "pre-specified" outcomes under the right developmental conditions. Thus, "specification" doesn't reside in the DNA alone, but in interactions.

(Schlichting & Pigliucci, 1995, 1998; Stearns, 1989; West-Eberhard, 2003).[2] But there are many kinds of phenotypes that are odd to think of in this way. For example, there are many kinds of human thoughts that are hard to think of as being orderable along a unidimensional continuum. If possible phenotypic outcomes in humans include thoughts about *Taxi Driver* and *Apocalypse Now*, what continuum could we put these on? Even if we found some unidimensional line on which they could be placed, how could this capture all the dimensions and nuances of the representation in a single number? And how could those be pre-specified?

They couldn't. These are examples of what is sometimes called "open" content, where there are many features or dimensions of the outcome that simply aren't specified and can't even be placed along any kind of sensible continuum, because there are many, many degrees of freedom. You might think this creates some kind of paradox, some inability to move forward through adaptive space because the reaction norm is unable to decide what phenotype to produce in response to the environment. However, the evolution of such systems doesn't appear to be hard at all. In fact, when it comes to systems that represent things, from perceptual systems to learning systems to decision-making systems, such open phenotypic outcomes are not just possible but the norm. And if natural selection couldn't select for reaction norms that have these kinds of phenotypes as outcomes, then representational systems could never evolve. So there have to be mapping functions that have phenotypes of this kind as output variables.

The way out of this conundrum, I suggest, is to realize that there are many, many functions beyond those that just map a curve from one continuous variable to another. What matters for natural selection is not whether reaction norms instantiate a simple two-dimensional curve, but whether the mapping functions are *causal*, even if they map in a complex way from one high-dimensional space to another. The relevant question is this: Is there a causal mechanism that maps between developmental circumstance and organismic phenotype, and are there variants of this causal mechanism that produce fitness-better or fitness-worse mappings in the environments the population experiences? If so, natural selection will shape the mechanism and therefore the reaction norm. This will happen whether or not all of its possible outcomes are pre-specified, in the sense of having previously been experienced and shaped by selection.

[2] The notion of classical reaction norms described here corresponds most closely with Schlichting & Pigliucci's (1995) notion of phenotypic modulation, in which phenotypic outcomes are continuous and proportional responses to environmental variation. The complex input-output mappings that result from contingent gene regulation, producing what they call "developmental conversion," begin to diverge from classic models as they become more like computational if-then rules and less like smooth curves.

In other words, classical two-dimensional mapping functions are just one subset of all possible mapping functions that take developmental circumstances as inputs and produce outcomes as a result. To see this, let's consider some examples from ordinary causation that show what open-ended causation might mean.

Car crashes are a kind of causal event that map from inputs to outputs: They take intact cars and turn them into wrecked cars. Suppose we studied car crashes and wanted to produce a computational description of the mapping function, or reaction norm, that car crashes instantiate. This could be a "lookup table" that tells us the exact details of how a wrecked car will look given the details of the intact car plus details of the crash circumstance (e.g., whether or not the collision was with another car or a train or a phone pole, details of the shape, form, mass, and composition of the other object, how fast each was travelling, the angle with respect to each other at contact, etc.). Or maybe it could be some set of equations that maps between input variables of interest and outcome variables of interest.

That would be one long lookup table. We might have more success with the equations approach, but even then, our equations could only hope to capture certain summary variables about the outcome, but not every detail of every specific possible outcome. Only an enormous set of equations could possibly produce an accurate detailed map of a crumpled fender or a photographic specification of the shatter pattern on the windshield as a function of the angle of the phone pole with respect to the car, how fast the car is rotating at contact, etc. In fact, the outcome is likely to be fractal in its detail, different every time down to the molecular level. Good luck finding the right equations to describe the input-output mapping other than in terms of some higher-level, abstract generalities.

Here's the point: In reality, the car doesn't need either a lookup table or a set of equations to produce its myriad of possible outcomes. The very high-dimensional input-output mapping function is actually determined by the causal properties of the car and what it collides with, and the "computations" occur on the timescale of the crash. From our perspective, there are many open dimensions that we can't ever hope to capture with a set of simple equations, but that doesn't prevent the input-output mapping from occurring.

What is the analogy, here, to developmental systems? Like variants of developmental machinery in organisms, different car designs instantiate different input-output mapping functions in collisions. Car engineers perform a crude analog of natural selection by varying car designs and putting them through crash tests. Of course, unlike natural selection, they can attempt to "guide" the variation using some basic principles, engineering equations and computer simulations to alter the design in desirable ways. But they still need to perform the tests to find their way up the design hill because of the inherently complex nature of the reaction norms their

designs instantiate. In the biological case of bodies and brains, populations create the real-world test cases that selection chooses between, engineering causal input-output mappings. Some areas of the mapping space will end up being "specified" to varying degrees of precision, meaning that selection will prune the causal pathways to reduce noise or random variation. Others will be more "open," meaning that selection will tolerate or perhaps even favor variation within some envelope or space of possibilities.

The specified and open dimensions or aspects of developmental outcomes need not be, and usually won't be, totally independent. In the case of cars, engineers might tolerate or even favor many degrees of open variation in the final shape of crushed cars, while constraining other degrees. For example, crash designs might allow lots of random deformation in certain dimensions of the car but very little in others, to deflect energy away from the driver. Developmental examples we've seen might include food concepts, where the nutritional evaluation system is selected to be reliably wired in particular ways to guide food search behavior while allowing for a wide or even open (unspecified) range of food types, and mental state concepts, where the inferential machinery is selected to develop similarly in everyone but the actual inferences made are open to circumstance. The variant and invariant parts are not necessarily separate dimensions or domains of the reaction norm, but instead, the variation is nested within the invariance: The crumpling of a fender after a crash is unpredictable and unspecified in its details, but it still falls within a certain parameter space, as does the food information represented in our brains.

Other examples of causal input-output mapping processes have outcomes that, while open, are less chaotic and more designed than those of car crashes. Record players have a design in which the vibrations of a diamond-tipped needle, caused by its movement along the shaped grooves of a spinning record, are amplified into audible sounds. Of course, this works because of a carefully human-engineered fit between records and record players. However, once that fit is in place, we can make records that can vary massively in content—the sounds recorded on the record—even though the input-output function that allows us to play them is highly specified. Obviously, records do not operate via a lookup table. The mapping function between groove shape and sounds that they instantiate can, in fact, be described by equations. But those equations wouldn't allow you to predict what sounds a record would make without seeing the record itself: Content isn't specified by the reaction norm, but a function for producing outputs from inputs is.

Another, more whimsical, example of a reaction norm that is even more open-ended—though not without limits and parameters—is garbage collection. You

have a can, you put stuff in there, you place it on the sidewalk, it ends up in the garbage dump. This is an example of what we might call a "carry-through" operation, where the function doesn't modify the inputs per se, but merely translates or moves them to a different location. In the brain, such operations might include simply moving neural information from one place to another without altering it. This is still a mapping function, but a very open-ended one: garbage in, garbage out. Of course, no such function can be *entirely* open-ended: Things must fit in the can, they must weigh less than the trash collector's maximum lifting capacity, and so on.

The point of these examples is this: For all reaction norms, there must be a causal process that maps between inputs and outputs, but that mapping function needn't be closed in the sense of having all inputs and outputs specified as *points* in some possibility space in the design of the norm. Those who designed record players presumably had no idea that the music of Black Sabbath would one day be played on them, and those who designed cameras—which map between patterns of light in the world and representations of those patterns on film or computer files—did not, and did not need to, represent each specific photograph their cameras would end up producing. Similarly, the evolutionary process, when it designed the mapping function between rod and cone cells in the retina and patterns of neural activity in the visual cortex, did not explore all possible objects you could ever look at—and yet the design works well for evolutionarily novel objects that were never "tried out" by the designing process. Contrast this with a lookup table, where if a value is not present in the table, the system doesn't know what to do.

This doesn't mean that computational formalisms couldn't be useful for modeling open reaction norms. We just need ones that, unlike functions that map from one space of small dimensionality onto another, can handle open content. For example, there are logic functions, such as if-then rules. Logic functions are reminiscent of garbage collection in that they accept highly variable content, such as propositions of the form "if P, then Q," and deliver outputs that carry through and recombine parts of the input, like "P" and "Q" (and they obey the principle of "garbage in, garbage out" in that if the inputs are nonsense, the outputs will be too). An example of a logic function is "modus ponens," a rule that algorithmically captures the principle that if P implies Q, and P is true, then Q is true. For example: If today is Thursday, then I will go to work (input 1); today is Thursday (input 2); therefore, I will go to work (output). The inputs can't be just anything; in particular, they must be propositions of form "if P, then Q" and "P." Any inputs that satisfy those minimal input criteria, however, will be processed. Like all functions, modus ponens relies on an inductive bet: namely that its output, "Q," is true if the inputs are, regardless of what they are (which, according to logicians, is a good bet). In essence, modus ponens instantiates

an inference rule, one that applies to inputs that must satisfy some important criteria but are otherwise open along many, many dimensions.[3]

Of course, what we want to understand here is not trash collection or abstract logic rules, but the causal stuff of organismic development. It is only when we think carefully about developmental mapping functions in causal terms that we can begin to make sense of words like "pre-specified" and formulations like "similar to the EEA along relevant dimensions." While we might, as scientists, describe reaction norms using algebraic functions, logic functions, or lookup tables, this is not (usually) how developmental systems actually do what they do. Instead, they are cause-effect systems made of biochemicals. In the case of brain development, we know that the outputs of reaction norms are systems that form representations, make inferences and decisions, and learn. The underlying mapping functions must, then, have many open dimensions. And yet they must also be extraordinarily well tuned in order to turn the chaos of experience into functioning brains, again and again.

Openness along many dimensions, then, is one feature we might expect many reaction norms of the brain to have. Related to this is the ability to handle specific instances of input that have never, as particular tokens or points in possibility space, been experienced by any member of the evolving lineage in the past. Imagine the input and output spaces of reaction norms as white spaces and place black dots in those areas of input and output space that individuals of the evolving lineage have actually experienced (i.e., where the design of the reaction norm has been tested by natural selection). What you will see for many systems, such as ones that develop our abilities to represent objects, faces, thoughts, and languages, is not one giant black plane (inputs) mapping to another giant black plane (outputs), but something more like one white plane peppered with dots mapping to another white plane peppered with dots. Many if not most of the representations we *could* end up developing never appear, such as sentences that could have been uttered but haven't been or faces that could have existed but haven't. Usually, however, novel sentences, faces, and objects bear many similarities to ones that *have* occurred in the past (because they are many regularities in the causal systems generating them; we'll turn to regularities in human-made environments presently).

This might lead us to expect the equivalent of "interpolation" and "extrapolation" in reaction norms. Interpolation involves estimating an unknown point by filling in a curve between two known points; in development, this would mean producing a development outcome that is novel in its details, but within the range of

[3] There is a debate in psychology about whether the mind reliably develops any general-purpose logical rules such as this, and the answer may well be no. I am not using modus ponens as an example of an *actual* reaction norm, but rather to illustrate that the mapping functions of reaction norms may often instantiate rules for operating on open content, rather than merely point-to-point mappings.

developmental outcomes previously experienced by the lineage. Similarly, extrapolation means extending a curve outside of the range of known points by following the best-guess trajectory that those points imply; development outside of the normal (experienced) range might often approximate extending the curve of the reaction norm from within the normal range. For example, the strange, artificially generated bodies and faces of films like *Avatar* or *Cars* are outside the range of previous human experience of bodies and faces, yet we can easily interpret their movements and expressions (extrapolation). Similarly, many sentences in English, or any other language, might never before have been uttered, and yet fall squarely within some range of variation that *has* occurred (interpolation). In development, we would expect reaction norms to generally be able to fill in the white spaces of possibility space that have not actually been visited, as points, very well when they are inside of the range of variation that has been experienced (interpolation), and perhaps fairly well even when they are outside the range of variation that has been experienced, as long as the system's inductive bets tend to hold there—or better yet, to the *degree* to which those inductive bets hold (extrapolation).[4]

All this suggests that what we might expect in brain development is a far cry from both the notions of pre-specification and a blank slate. Some *aspects* of reaction norms might be curve-like, or lookup-table-like, or if-then-statement-like, but these do not describe the sum total of input-output mapping any more than the shape of a record-player needle tells us what sounds will come out of the speaker in the absence of knowledge of the record it's playing. It is only by thinking about what happens when needle meets record, or mind meets world, that we can begin to understand phenomena such as the handling of environmental variation and novelty.

---

Let us now turn from metaphors to real developmental systems and ask how open-ended systems might evolve. Keep in mind that there are virtually no cases of developmental systems in the brain for which we know all or even most details of how the reaction norms occur causally. And the approach I'm advocating here, which traffics in concepts of design features for variances and invariances, won't necessarily tell

[4] Some methods for modeling development assume that the evolutionary process has explored the entirely possibility space of development. Dynamic programming, for example, uses a technique known as backward induction to find the optimal path through developmental space (Frankenhuis et al., 2013b; Mangel & Clark, 1988). While this allows us to make predictions about how development might optimally occur—similar to the use of Bayes's rule to make predictions about optimal learning and inference—it doesn't imply that selection actually *has* tried out every possible pathway.

us all of those details. Instead, what I'd like to do is think about such systems at an abstract level in terms of fit. What are the fit relationships that evolve in such developmental systems, and what happens when you place them in the world?

An instructive example to begin with is one that is well known in the developmental literature: the development of celestial navigation abilities in a small migratory bird called the indigo bunting. The design features of this system were first explored in detail by biologist Stephen Emlen using an ingenious set of experiments (Emlen, 1975). It had been known that buntings migrated long distances with impressive accuracy, and that they appeared to use the stars in the night sky to do so. By raising buntings from birth in a planetarium, Emlen could control the input the buntings' developing brains received: He could project different spatiotemporal patterns of stars onto the planetarium ceiling during the buntings' childhood and then measure in which directions they flew when it was time to migrate. What Emlen found was that the developmental systems of bunting chicks are not designed to rely on any *particular* star or set of stars for orientation. Instead, they exploit a constant property of the night sky: the fact that stars rotate around a central axis of rotation (which defines an apparent central point of rotation in the sky when light from stars is projected onto the retina). Using his planetarium, Emlen was able to choose an arbitrary point in the night sky and make it the center of rotation, causing the projection to rotate around that point. He showed that birds raised in these conditions would learn that this arbitrary point was the center of celestial rotation, and use this when deciding which way to take off in the real world. A variety of other phenomena he discovered made sense in light of this, as well: For example, removing single constellations had no effect on the birds' flight (as one might expect if they were programmed to rely on particular stars), but blacking out all the stars near the center of rotation confused the birds, whereas leaving only the stars near the center left their navigational decisions relatively undisturbed.

There are several points to take from this example. One, most obviously, is that developmental systems can be and often are designed to embody assumptions about the variances and invariances they will encounter in the world. As I mentioned, terms like experience-expectant learning, prepared learning, and guided learning are used to describe cases such as this, because the developmental systems involved systematically use certain properties in their learning (Barrett & Broesch, 2012; Gould & Marler, 1987; Greenough et al., 1987; Seligman, 1970). However, on the view I'm endorsing here, prepared learning shouldn't really be seen as a special category of learning, since *all* learning mechanisms can only work if they embody assumptions about the structure of the input. In fact, any system that makes inductive bets, as I'm arguing that all adaptations do, can be seen as being prepared to encounter and deal with some dimensions in the world that are variable, and some that are not.

Clearly, there is domain specificity in the bunting system, and a high degree of specialization. Not all learning designs could pick up on celestial rotation around an axis. The algorithm or mapping function that turns the observation of slowly moving stars into adult migratory decisions must be a specialized one indeed, and one that many species, including humans, don't have. And yet, its specialized algorithms seem to be able to handle open content: namely, whatever pattern of stars is rotating around a central point. Indeed, it might even be able to handle evolutionarily novel content. Of course, we can't know for sure, but buntings transported to a planet similar to ours in another galaxy or our own planet far in the future, with an entirely different set of stars rotating around a central point, might be able to do just fine (indeed, the exact configuration of stars in the sky has subtly changed over the course of buntings' evolutionary history). If the buntings responded adaptively, this would be because the novel inputs would satisfy the inductive bets of the buntings' navigational systems. An axis of rotation plus stars rotating around it might be required, but the stars needn't necessarily be *our* stars.

There are several concepts that are useful for thinking about developmental outcomes for open-ended systems such as these: types, tokens, and adaptive targets. The adaptive target of a developmental system is the *type* of outcome that the system has been selected to produce (Barrett, 2006; Tooby & Cosmides, 1992). One might think of this as a summary description—usually at a rather abstract level—of the common features of the token developmental outcomes that the system has produced in the past that have increased fitness. The tokens, of course, will vary in many details, and it is not that these specific details haven't been important to fitness. For any given bunting, for example, it is a specific pattern of stars—a token of all possible star patterns—that allows it to find its way home. However, that specific pattern is not engineered into the reaction norm itself. Instead, the reaction norm is designed to acquire it. Thus, the particular pattern of stars that a given bunting might know is a token of the type of developmental outcome that the bunting's developmental system has been selected to produce (i.e., a star map with the center of rotation calculated). We can call this type the adaptive target that the developmental system has been designed to hit. And when development occurs in environments that match the EEA of the reaction norm along the proper dimensions, the adaptive target tends to be reached—what is known as *reliable development* (Barrett & Behne, 2005; Tooby & Cosmides, 1992).

Thinking about reaction norms, adaptive targets, and reliable development is quite different from thinking about innateness in important ways. If innateness means "present at birth," buntings' navigational skills are not innate. If innateness means "develops without the influence of experience," their navigational skills are also not innate: no experience, no navigation. One can, of course, conceptualize the

reaction norm itself as innate. But even on this view, no specific phenotypic outcome itself is innate or specified by the genes alone.

Certain things about this perspective should give us pause when thinking about how to interpret empirical findings about development. Sometimes, for example, developmental outcomes will have the *appearance of innateness* even when they are neither present at birth nor uninfluenced by environment. If a reaction norm uses environmental information to shape the phenotype and that information is present in all or most human environments, the phenotypic outcome will be the same everywhere. And this can be true even if those reaction norms would have developed differently if placed in different environments (Barrett, 2006).

In the northern hemisphere of earth, for example, the night sky rotates around Polaris, the pole star. Thus, all buntings raised in the northern hemisphere will use Polaris as the center of rotation and will have it represented as such in their star maps. Does this mean that the use of Polaris is innate? No. There is merely the appearance of innateness, because all buntings happen to develop in a world where Polaris is the center of rotation. Place them in a different environment—even an evolutionarily novel one, like the southern hemisphere or perhaps even another planet—and they would (perhaps) do just fine using a different set of stars.

What I'm calling the problem of the "appearance of innateness" is one that we must be mindful of when thinking about humans too. If the view of reaction norms I'm presenting here is right, much of development will systematically use information that is not coded in the genes in order to guide development—even when, and sometimes because, that information is entirely constant in the world. For example, all human children develop an intuitive understanding of gravity early in childhood (Baillargeon & Hanko-Summers, 1990). The early and apparently universal development of skills such as this, as well as other aspects of our intuitive physics, is sometimes taken as evidence that these are innate. They are certainly reliably developing, given the terrestrial environments that all human children inhabit. But what would happen if you raised a baby in the zero-gravity environment of outer space? Would the sense of gravity develop in the same way and at the same age as the most intuitive sense of "innate" implies? We don't know. But crucially, even if it didn't, that doesn't mean that natural selection hasn't favored a capacity to understand gravity; indeed, a gravity-acquiring reaction norm is a plausible biological proposal.

Gravity might not seem a huge problem because gravity-specific adaptations seem biologically plausible. But slightly more absurd cases show that the inference of innateness from universality alone can lead to false conclusions. Consider concepts that might develop "universally" in all children, or nearly so, in today's hyper-globalized world. At the current moment, these might include the concept of spoons, the concept of television, or even the concept of Batman. This universality, of

course, presumably comes from some set of evolved reaction norms interacting with current (but not ancestral) environmental universals. We wouldn't want to conclude from this that the concept of Batman is innate or that there were adaptations specifically for reliably developing a concept of Batman (Batman might be a token of a type of concept we've been selected to acquire, but that's precisely the point). To make inferences about adaptations, we need more specific hypotheses about the design and function of the underlying reaction norms. Unfortunately, however, many people consider universals and variation alone as evidence for or against evolutionary accounts, such that developmental invariances across individuals and environments count as evidence for evolutionary accounts of development, and variance counts against them. This clearly needs to be rethought. It's not that variance and invariance aren't important sources of evidence; it's that they only make sense in light of hypotheses about evolved reaction norms.

---

Let us now turn to the relationship between open reaction norms and novelty: the ability to react to environments and inputs that have not appeared in the past and to produce outcomes that selection has not specifically selected for. Novelty, and the ability of evolved systems to handle it adaptively, seems to hold an air of mystery for many who contemplate evolution and how it works. Given that adaptations evolved in the past, how is it that they can work—and apparently work well, in some cases—in the present and in the future? How can they deal with new situations, ones that the ancestral versions of the adaptive design never encountered?

Humans, we are endlessly reminded, are masters of novelty. It is true, as I mentioned at the beginning of this chapter, that we routinely do, make, say, and think things that have never before been done, made, said, or thought in the evolutionary history of our species. And there is widespread skepticism that specialized adaptations that evolved in the past can have anything to do with explaining these aspects of the present that are new and historically unprecedented (Chiappe & McDonald, 2005). How can adaptations that evolved to solve problems in the past have anything to do with us thinking about movies or computers or forks, since these are all evolutionarily novel?

The answer, I have been suggesting, lies in the inductive bets that evolved mechanisms make, and the fit between these bets and the world. Whatever truth there is in this statement, however, relies on us being very precise in how we think about and define novelty. The most appealing and intuitive definition of novelty highlights exactly and only what is new, what has never occurred before. And if we define novelty in this way, one thing is clear: The idea of an adaptation to novelty is an oxymoron, a

contradiction in terms. Adaptation to pure novelty is impossible because if nothing about a novel input or event bears any resemblance to past ones, then there cannot have been selection for it. If we choose to define novelty in this way, then any adaptive handling of novelty can only be a matter of pure coincidence.

However, the examples that are often given of evolutionarily novel circumstances, events, inputs, and activities that humans engage in—things like reading, writing, driving cars, and playing chess—suggest that this extreme definition of novelty as bearing *no* resemblance to the past is probably not right. In fact, on the massively multidimensional view of environments and information that I have been advocating, every event, object, input, and activity has *aspects* that are novel, and aspects that are not. This was true even of the situations in the past that shaped the design of present adaptations: Every event had some unique and idiosyncratic properties that would never and will never reoccur. But natural selection is a statistical process that is only possible if there are patterns over time, statistical signals of resemblance or similarity between events. We can see this illustrated in the bunting case, where the developmental system has averaged over the details of the individual stars to narrow in on a statistical constant of the night sky, namely, rotation around a central point. What this means is that if buntings are placed in an environment of novel constellations, their developmental system will operate adaptively as long as those constellations also rotate around a central point. In this regard and perhaps a few others (e.g., the fact that stars are points of light within a certain size and wavelength envelope, distributed against a background of black), the new constellations are not novel. They continue to satisfy the inductive bets engineered into the system, and not by coincidence.

Thus, given that any situation or stimulus that an organism experiences has many properties or aspects, some of these may be genuinely evolutionarily novel and some may not be. We can think of this in terms of aspectual novelty and aspectual ancestrality. Intuitively, it seems as if many if not most inputs to the mind in modern environments probably carry some mix of aspectually novel and aspectually ancestral features. This way of thinking about novelty—not as an all-or-none property, but as a matter of aspects of resemblance to past environments in ways that satisfy an adaptation's inductive bets—can actually help us to predict when systems can be expected to respond adaptively to novelty, and when not. Let us think about this with respect to another type of largely (but not entirely) open content that our brains have evolved to handle: food.

In chapter 3, I reviewed the kinds of decisions that animals routinely make about food: Eat this now? How much to eat? When to stop eating and start searching for something else? The fact that animals make decisions such as these—and in the case of omnivores with broad diet breadths like humans and rats, with a high degree of

openness for different kinds of foods—suggests systems designed to handle inputs that are variable in many ways, yet that obey a certain logic. Our food evaluation systems must categorize food into categories, associate particular nutrient values with each category, represent how these foods are distributed spatially across the landscape, and combine these representations in particular ways—ways that sometimes reflect complex calculations—to make decisions.

What must the reaction norms that build such food evaluation systems into animal brains be like? In this case, the adaptive target first includes the representation types required for the system to work: For example, it must be able to estimate the nutrient content of different foods, categorize them based on perceptual cues like shape, color, and smell, and represent how they are distributed spatially. Second, it includes the decision rules necessary to process the information in those representations to produce behavioral decisions. As elsewhere, the adaptive target in this case includes some dimensions that are specified and others that are open. For example, the food evaluation system must have rules that tell the organism things like how much time to spend looking for food in a patch as a function of the expected value of food to be found there, how to decide which foods to search for as a function of tradeoffs between nutritional value and difficulty of obtaining the food, and so on. And yet one can imagine a very open-ended foraging system in which little or nothing about *specific* food types needs to be specified. In humans, for example, the same system will work fine whether you're eating nuts, berries, worms, rabbits, or pizza, as long as your digestive system contains nutrient detectors that are hooked up in the right way to the brain, categorization systems that chunk them properly, and so on. In other words, the adaptive target of the system includes *types* of representations and rules for processing them. The content of the tokens that end up being represented is open along many dimensions—it can range from hummus to donuts—but it is not limitless content (it must be food).[5]

We don't yet know all the details of how such systems develop, but they must involve specialized procedures for wiring some systems to others—nutrient detectors

[5] Note here, as for the other specialized systems we've talked about, that not everything about food processing need be unique to food. For example, the basic logic of optimal foraging has been found to hold in foraging for nonfood resources, searching for information on the Internet, and even in memory (Hutchinson et al., 2008; Pirolli & Card, 1999; Wilke & Barrett, 2009). Indeed, Hills (2006) has suggested that foraging is the original source domain for what he calls goal-directed cognition, or cognition involving the search for rewards. On this view, ancient foraging mechanisms for adjusting behavior based on nutrient wins and losses, mediated by dopamine and other biochemical signaling pathways, have been modified for other, nonfood-related aspects of cognition (e.g., risk-taking). This doesn't, of course, imply that food-specific aspects of cognition have been lost.

in the bloodstream to computational devices in the brain, and object categorization systems in the visual system that match visual appearance to estimates of nutrient content, for example—and specialized learning procedures for adjusting these connections and producing the right representations and decisions, based on the information passing from one system to another. Critically, these developmental systems must make massive use of specialized division of labor, with some systems doing some calculations (nutrient evaluation) and other systems doing others (learning how to group types of visual objects, learning spatial distribution of foods in the environment, and adjusting foraging decisions as a function of these). And to hit the adaptive target, the developmental system is likely to use feedback among these specialized systems as they develop. Now let us consider the relationship between selection in the past and the design of the system in the present that will allow it to sometimes respond adaptively to novel inputs, and sometimes not.

The food evaluation systems of any species evolved because certain generalizations held across objects and situations—food items, their nutrient contents, and their distributions in space and time—in ancestral environments. And the details of any given species' food evaluation system, as for any adaptation, will be a mix of species-specific characteristics and ones that are more phylogenetically ancient. Our system, for example, was handed down through ancestors who started out eating insects many millions of years ago, then transitioned to plant foods (especially fruits), then to more varied diets including, relatively recently, meat (Stanford & Bunn, 2001), and very recently, the products of agriculture (Perry et al., 2007). At each stage, the inherited design was both modified in certain ways and conserved in others. But when thinking about the type-level design features of the system, what matters is that the set of tokens that shaped the design of the system is a giant fuzzy cloud of *actual* food items, actual environments. One individual was eating some berries and was too slow to move to the next bush and he died, and his design variant was lost. Another was eating bananas and decided to move along at just the right time, and got a little fatter than his neighbor, and had more offspring. The design that we see now is the result of these events added up over time: natural selection.

What, then, is the type that the food evaluation system was designed to handle? The obvious answer would seem to be "food," and that is indeed a description that casts a wide-enough net to encompass all of the tokens that shaped the design of the system. But we must not lose sight of the fact that what we're describing as a type was in fact a cloud of tokens spread across ancestral time and space. In case you think this kind of technical distinction can't possibly matter—you would prefer to just say that "food" is the proper domain of the mechanism and be done with it—consider the following: A squirrel encounters a chocolate bar and decides to eat it. His food evaluation system digests the bar, assesses it as high in nutrients, and causes the squirrel to return to that location

later in search of more. Is the system doing what it was designed to do? How about when a bear learns that trash cans in public parks are good foraging locations and develops a routine in which he goes from can to can, adjusting his routine to visit the most rewarding cans more frequently? Is his foraging system responsible, and is it working properly?

I think most of us would say yes, the food system is doing just what it was selected to do in those cases. This is not to say it can't make mistakes or result in maladaptive outcomes, like the squirrel or the bear getting tooth decay, or worse. What's important is that it seems reasonable to say that even when the bear is returning to a campsite because of half-eaten bags of cookies and hot dogs, its food system is doing what it was selected to do—even though cookies and hot dogs were not in the EEA of the bear's foraging mechanisms. What is doing the causal work here is the resemblance of cookies and hot dogs (and what they are made of) to things that actually *were* in the EEA of the bear's foraging mechanisms. It won't do to say something like "the proper domain of the system is food, and cookies are food, so they are part of the proper domain." Calling the proper domain "food" is a summary description that we, as observers, are making, but there is no doubt that cookies were *not* encountered by ancestral bears; they are evolutionarily novel. The reason that cookies interact with the design of the optimal foraging system to produce adaptive outcomes is because the cookies have properties that match those of the tokens that drove the evolution of the system along relevant dimensions. For example, they contain nutrients, and they are distributed in special patches whose timing and location have some predictability. And these objects satisfy the input criteria of the optimal foraging system, causing them to be processed: The bear smelled them, ate them, remembered what they looked and smelled like, and associated the good feelings produced by nutrient detectors in his bloodstream with things that look and smell like that. And this was not a coincidence, unless you think that it's a coincidence that cookies, like berries, have nutrient content and are distributed in space and time in semi-predictable ways.

What's going on here, of course, is that the design of the bear's foraging system is making inductive bets, and those bets are satisfied for trash cans, as well as for berry bushes. In the language introduced in chapter 1, they satisfy the inductive bet's felicity conditions. I mentioned that there is a large philosophical literature on induction, part of which asks what "justifies" inductions: When, if ever, can we be sure that an inductive generalization is true? The short answer is that we can't ever be completely certain. This is because what we're doing is predicting the properties of some as-yet-unseen tokens (objects, events, experiences, arrangements of things in space and time) based on tokens we've already seen. We're predicting the properties of novel tokens based on experienced ones. The only thing that "guarantees" that our inductive generalization will apply to the novel tokens is that they do, in fact, resemble the tokens we've seen along the relevant dimensions. There is no general rule for making safe inductions

other than this: Make them along dimensions where regularities actually exist in the world. This is why the process of natural selection in the past is necessary for systems to make good inductive bets now. Those bets are based on the number-crunching process of natural selection meeting the hard reality of the world over periods of time and space. It is also, at the same time, the only reason why the inductive bets engineered into evolved mechanisms work well in the present in a systematic way.

These considerations lead to what we can call the principle of *felicitous novelty*: An adaptation interacting with novel stimuli will produce adaptive outcomes when those novel stimuli satisfy the fit relationships that the adaptation evolved to exploit (i.e., its felicity conditions). It shouldn't be a particular surprise that evolved inductive bets often work out felicitously in novel situations, because there are many reasons why the regularities that made bets work in the past will sometimes extend to new situations. For example, brownies are solid objects made on purpose by humans out of ingredients that are (mostly) nutritious to us, an example of a culturally evolved artifact designed to fit the evolved designs of our minds and bodies. The felicity, in this case, is by design: a combination of culturally evolved design and human intentions. Of course, there are also cases where the conditions for handling adaptive novelty will be systematically infelicitous. For example, optical illusions are designed to trick our visual system by systematically violating the principles on which their inductive bets rely, and nest parasites have been designed, by natural selection, to trick the inductive bets of the poor host birds in whose nests they lay their eggs.[6] And of course, there are many cases of novelty that will be felicitous or infelicitous by chance.

Rather than assume that evolved mechanisms simply *can't* handle novelty, then, we need to develop specific hypotheses about the inductive bets that mechanisms instantiate, which will in turn tell us something about their *robustness* (Gigerenzer et al., 1999; Kitano, 2004). Robustness is a concept from biology, psychology, and information science that refers to the ability of a mechanism or system to produce adaptive outcomes under varying, novel, or uncertain conditions. In the case of evolved psychological mechanisms, rather than considering them as destined to fail in novel environments, we should ask: What are their adaptive felicity conditions, the sets of conditions under which they will operate properly? And on the flip side, under what conditions are they likely, or guaranteed, to fail?

---

[6] In the domain of food, such tricks exist as well. For example, artificial sweeteners have been designed by chemists to trick our sugar detectors. Interestingly, this is something that people want to be tricked about, though whether artificial sweeteners actually increase or decrease fitness is an open question.

Let us consider one more example of an open developmental system, which has as its adaptive target phenotypes that are capable of representing a vast diversity of information: the system that builds our capacities for representing the contents of others' minds.

As I discussed in chapter 5, the ability to represent and make inferences about others' internal states is a paradigm case of a higher-level, complex, multifaceted, flexible ability. It is higher level because the representations and inferences are far removed from the initial stages of perception. They are conceptual, involving things like beliefs and intentions, and in order to build them, many post-perceptual inferences are required. The ability is complex in the sense that it undoubtedly involves many interacting mechanisms and systems. It is flexible both in the sense of variable content—many, many kinds of mental states can be entertained, from "George prefers full-bodied wines" to "Scientologists believe that humans are descended from a race of clam-like aliens"—and in the sense that mental state inferences are highly context-sensitive and can be influenced by many different kinds of information in the mind.

These properties present a challenge to many peoples' ideas of what a specialized, domain-specific system must be like. Theory of mind isn't just a reflex, mapping from some perceptual cue in the world to a behavioral reaction. Indeed, as I've mentioned, many of the representational structures that are part of the adaptive target, like concepts of belief and desire, are far removed from perception and are not marked in the world by simple cues like the pleasant smell that indexes the presence of brownies. In fact, most psychologists believe that if things like beliefs and desires exist in brains, it is not as objects or things per se but as higher-level abstractions in the form of patterns of information and neural connections (Dennett, 1987). How could a developmental system evolve that has these things as its adaptive targets? And given the diversity of beliefs, desires, and other mental states in the minds of others, what if any are the commonalities among them, if any, that natural selection could use to engineer inductive bets?

Let's focus on beliefs. The first thing to note is that beliefs are not *entirely* without perceptual correlates. If there were no perceptual cues to beliefs at all, we would have no way of inferring what others believe. Consider, for example, the bank of gaze-detecting neurons in zebra brains that I mentioned in chapter 5: These cells fire when the zebra has been looked at and remain active for some time afterward. Crudely, this firing pattern indexes having been looked at. If there were a way to couple the information states of these neurons with representations of individuals—like the individual lion who just looked at me—then we would have a system that could register, or index, the fact that the lion *knows* where I am. Of course, this is far from capturing the sum total of human belief-tracking abilities. But it does give us an angle into thinking about the kinds of inductive bets a belief-tracking system

might instantiate, and what kinds of variances and invariances a developmental system might use in building such a system.

Consider the resemblance between this crude system of belief tracking and Alan Leslie's proposal about M-representations (Leslie, 1987, 1994). Recall that Leslie suggested that our belief representations come in a particular, composite representational format that can be summarized as [agent]—[attitude]—[proposition]. This is a composite representation in that it links three other kinds of representations in a particular relationship. In the zebra case, the agent would be the lion; the attitude would be the attitude of seeing or knowing; and the proposition—which, remember, simply means a representation of a state of affairs, and could come in the form of a spatial state of affairs, a temporal state of affairs, or a conceptual state of affairs—would be a representation of the zebra's location (or alternately, the zebra's presence or existence).

Such a system, if it is indeed an approximately correct description of how belief representations work, has two interesting properties. One is that it involves a computational division of labor, with different parts of the system handling part of the belief-computing problem and collaborating to produce a composite representation—a topic to which we'll return in chapter 11. Second, the system at least potentially can make use of some parts, or representational systems, that predate the evolution of belief-tracking abilities per se: for example, the capacity to detect and individuate agents, like the lion, and the capacity to represent states of affairs, as in mental maps of the environment. This implies that engineering a belief-tracking system wouldn't require engineering all components of that system from scratch, but instead modifying and adding to systems that were already there, hill-climbing from a point on the adaptive landscape where many of the requisite abilities were already present. As we discussed in chapters 1 and 2, this ability to make use of capacities that were already in place is likely to be an evolutionary virtue. It means that at some point, the stage might have been set for the evolution of theory of mind in the sense that mutations or variants in brain wiring that happened to connect the preexisting systems in fortuitous ways could have given an initial bump in fitness when they first appeared. In other words, the evolution of belief tracking could have involved a substantial amount of path dependence: Not only would it not be equally evolvable in all species, but even those species poised at the requisite place on the adaptive landscape might not all experience the requisite tweak in brain wiring that would open the pathway up the adaptive hill toward belief tracking.

Obviously, much here depends on the details. What do I mean by a fortuitous change in brain wiring, and how is this anything other than positing a magical evolutionary leap?

If Leslie's or any functionally similar account of belief tracking is right, then belief tracking requires "noticing," or representing in neural form, the relationship between several kinds of things in the world. First, you have to be able to represent states of affairs, like "I am here right now," or, perhaps more formally, to represent one's own location in an internal mental map of the landscape. Second, you have to be able to register other agents' attentional states, such as where they are looking. Finally, you have to be able to form a representational link that encodes the right kind of relationship between the attentional state and the state of affairs: what Leslie is calling the agent's attitude toward the state of affairs, and what in plain language we might call concepts of belief, knowledge, liking, and so on.

In the scenario I'm suggesting, let's assume that two of these abilities are already present: the capacity to track others' attentional states and the capacity to form representations of states of affairs, such as one's location in the spatial environment.[7] The evolutionary innovation required, then, would be the ability to begin to represent particular kinds of relationships between agents' attentional states and states of affairs. In particular, these would have to encode something like the causal consequences of an agent having attended to a particular state of affairs (e.g., your location) in terms of how that agent will later act.

What is needed, then, is the ability to notice this relationship between an agent's attention, particular states of the world, and how the agent's future behavior depends on those states (for example, he last saw me in location X, and even though I'm now at location Y, he's looking for me at location X). And here, we might invoke an ability that we know that the neural networks that make up our brains are very good at: detecting contingencies (Watson, 1985). This property of neural systems has been studied extensively since the early work of pioneering neuroscientists like Donald Hebb, who proposed that neurons that "fire together, wire together"—an inductive bet that instantiates contingency detection (Hebb, 1949). Since Hebb's time, the details of how neural networks encode contingency have been much refined, but his point, that neurons are well designed to register spatiotemporal coincidences in the activity of other systems, still stands (Gallistel & King, 2009). What this means is that if banks of neurons were to appear that connected systems for tracking other

[7] On this account, the scope of belief contents we can represent is directly dependent on the scope of *states of affairs* we can represent. This means that the content of theory of mind representations depends on the representational capacities of many systems that are outside of theory of mind per se, and many of which may predate it (e.g., our abilities to represent space, time, causation, etc.). Again, this is a good example of specialization through interaction, which contradicts the standard view of modular systems as isolated and sui generis (see chapter 11).

agents' attentional states and one's own representations of states of affairs—and if those neurons possessed a memory of the relations between these that could be correlated later with the agent's future behavior—then the stage would be set for a system that could notice, or respond to, the causal relationship between agents' attention, states of affairs, and the agent's future behavior contingent on those states of affairs.

This is hypothetical, of course, a conjecture. But note that there is nothing magical about it. The relevant evolutionary step, the initial additivist change, would be the appearance of new neural wiring connections between systems. At first, this could have been due to random variation in neural wiring, and the neurons in question would not yet be specialized for representing "attitudes" in the fully formed Lesliean sense of concepts of belief, desire, etc. But the relevant mutation, or change, would have occurred that would open up a pathway on the adaptive landscape that was not previously accessible. To be preserved, the new wiring pattern would have to have immediate fitness benefits in the sense of being capable of registering, at least crudely, some relevant contingencies between the connected systems. However, it would be through subsequent variation in design and retention of the fitness-better variants that evolution of a truly specialized system for belief tracking would occur. This could include selection for mechanisms prepared to detect specific contingency signatures diagnostic of beliefs, desires, intentions, and the other mental state concepts we are able to develop, leading to the reliable development of these concepts in most children (the scenario I've depicted here doesn't rule out the ability to develop novel mental state concepts as well).

What does this evolutionary scenario, or a scenario like it, imply for the design of the developmental systems that build theory of mind? Synchrony and temporal contingency of neural firing are likely to be important shapers of brain development (Uhlhaas et al., 2010). And the alteration of wiring patterns between brain systems is likely to be an important source of evolutionary change (Striedter, 2005). Both of these fit with the scenario I am giving here. If it's right, then what matters for reliable development of belief tracking are that the developmental systems that build the appropriate components of the system are in place (agent tracking, attention tracking, world-state tracking), as well as patterns of wiring among them that allow the relevant contingencies to be reliably and consistently detected during childhood.

We could call this developmental system a prepared, experience-expectant learning system that bets on certain contingencies in the world in order to wire itself properly. But whatever we call it, one thing is certainly true: In order to understand how it works, the question is not whether plasticity *or* adaptation by natural selection explains the outcomes it produces. It's whether the right kinds of plastic designs,

designs composed of parts with specific functions, have been shaped by a process of variation and selective retention over evolutionary time.

---

The notion of open-ended reaction norms—norms that instantiate functions in many-dimensional phenotype spaces and that have different shapes (and sometimes no shape at all) along different dimensions of the space—requires rethinking of some currently popular ideas about mental evolution.

First, the idea that natural selection needs to specify phenotypic outcomes in order to shape the design of phenotypes is misleading. It is an appealing idea if one thinks in terms of classical reaction norms, but not for open ones. In order to operate, natural selection does not need to have "seen" every possible outcome that its adaptations can produce. It does not create a one-to-one lookup table or mapping function that maps between every possible input and output, even though we can sometimes model reaction norms in this way. Instead, the mapping functions that selection creates are causal engines. Selection engineers into them inductive bets that tended to have been statistically fitness-good in the cases that have occurred in the past, but that only *will* be fitness-good in new cases when those cases satisfy the system's inductive bets.

This has empirical implications for how we do science. Because organisms are complex objects in a very high-dimensional space, every measurement we make is a kind of transect, or slice, through that object. It will therefore be possible to create what seem to be like closed reaction norms—point-to-point mapping or closed curves—by choosing to make transects or slices of very low dimensionality: for example, looking at the relationships among genes, environment, and numerical scores on an aptitude test. This is entirely legitimate and is, of necessity, how much of biology, psychology, and neuroscience proceeds. We often pick transects through the organismal space that will give us readouts along dimensions we can measure, such as nutrition, height, or performance on a laboratory task. But we mustn't forget that each such transect tells us only about that dimension, and not the ones we haven't measured.

A related point is that we need to think carefully about the levels of abstraction or precision at which we are describing developmental systems and outcomes and how these are instantiated in physical stuff, paying attention to distinctions between types (more abstract) and tokens (more precise). We can say that brains execute Bayesian updating, or do time-series analyses, or represent mental states, but how is this instantiated in the material system of genes, neurons, and neurotransmitters? When describing physical, biological systems, we have at our disposal descriptions

and models at a variety of levels, all equally legitimate to the degree that they capture biological reality—but we must be mindful of biological realism when picking our models. In most cases, it makes little sense to think of natural selection as specifying point outcomes of phenotypes in a massively multidimensional space; phenotypes are not preformed in this sense (Lickliter & Honeycutt, 2003; Smith & Thelen, 2003). And classical reaction norms, to the degree that they specify closed spaces of point-to-point mappings, are unlikely candidates to account for all, or even most, of brain development. However, the alternatives that have been proposed, to the extent that they are even *simpler* (i.e., a small number of learning rules to account for all of development), may well be going in the wrong direction. In all likelihood, we'll need a larger number of more complex reaction norms, rather than a smaller number of simpler ones, to account for how the full complexity of adult brains gets built from a single fertilized egg.

With regard to how brains handle evolutionary novelty, the currently fashionable dual-systems solution is to carve the brain's mechanisms into two categories: rigid (specialized) and flexible (non-specialized). Instead, I'd suggest that we take seriously the idea that different systems in the mind instantiate different inductive bets, but no system can function without any design features at all. If specialization is achieved through interaction of many parts, some older and some newer, then the way these parts will respond to novelty is far from obvious. It requires careful study of the felicity conditions of these inductive bets, and what happens when you mix together a bunch of them. It's surely *possible* that one box of mind parts is responsible for how we react rigidly and innately to evolutionarily old stuff, and another box of mind parts is responsible for how we react fluidly and contingently to new stuff. However, I see no good reason, biologically, to expect this to be the case. Instead, novelty often seems to interact felicitously with our evolved adaptations. Rather than just invent a word, plasticity, to describe this, we ought to try to find out why.

Two final points: First, as I've mentioned, evolutionary novelty is often infelicitous. Indeed, in some circles, cases of evolutionary disequilibrium or adaptive lag, such as our (apparently) maladaptive cravings for salty and fatty foods or the (apparently) predictably irrational features of our everyday decision-making, are favorite cocktail party openers for discussions of evolutionary psychology. Far from arguing that brains *always do* the right thing when placed in novel environments, I'm arguing minimally that they *can*, and I'm arguing somewhat more strongly that they *often do*. Moreover, I'm suggesting that by thinking about design features and inductive bets—as opposed to, for example, just plasticity per se—we'll be able to make progress in understanding both when adaptations do well in novel circumstances, and when they go awry.

Second, when evolutionarily novel situations *are* felicitous, it's often, or perhaps even mostly, not a coincidence. The world is a chaotic place, so if evolved mechanisms properly handle things they've never been exposed to before, there's probably a reason. One very important kind of reason is that the world is structured by many, many processes that continue to hold true today just as they did a million years ago. Another very important kind of reason is that the world responds to *us*. This is the topic to which we will now turn.

PART IV

# Culture

# 8

## MOVING TARGETS

It's a fact of the world that environments vary. There is variation both between environments and within them. Drive your car across the United States and you will see environments change before your eyes. Stop your car at any point and look around and you will see diversity, even on the bleakest median strip in the middle of the highway. A seed landing and growing into a plant at any point on this journey would have to deal with different circumstances—different soil, different sunlight, different neighbors—than a seed landing somewhere else.

It's clear that if adaptation is to happen at all, it must be able to deal with such variation. Moreover, the adaptations we see *have* dealt with it: Evolution casts organisms across environments like seeds across a patchwork landscape, and every living organism embodies a design that has managed to survive all the hands it was dealt, passing through a long line of individuals, circumstances, and environments and managing to reproduce itself successfully at every step. We have seen how natural selection acts as a giant number-cruncher, engineering designs that are fitted to the breadth and depth of variation the population has experienced over time, accommodating to its nooks and crannies and statistical shapes. Some of these designs are static and some—probably most—are plastic in some way, prepared to react appropriately to the circumstances they encounter.

But evolutionists have discovered that some variation is trickier to prepare for than others, because the variation itself varies. There are two well-known examples in biology: predators and pathogens. Predators, of course, are organisms that try to catch and eat prey, and they are designed by evolution to do so. Pathogens are organisms that infect or attack their hosts in some way, using them as food, reproductive vessels, or both—and in many ways, they are much more problematic than predators because of their ubiquity in our environments and their relatively shorter generation times, allowing many generations of natural selection to occur in the lifetime of a single host. Both predators and pathogens are elements of the environment to which species must adapt or die—and yet they are features of the environment that *also* change, adapting right back. As prey or host species evolve defenses against predators

or pathogens, the latter counter-evolve better and better means of defeating those defenses. This process has been likened to an arms race. Biologist Leigh Van Valen dubbed it the "Red Queen" phenomenon, after the character in *Through the Looking Glass* who must run faster and faster just to stay in the same place (Van Valen, 1973; Ridley, 1995).

What's interesting about these situations is that despite the fact that the environment is always changing, and even changing in a manner specifically designed to track and thwart counter-adaptations, adaptation nevertheless occurs: Adaptations evolve. As you might expect, some of these adaptations, or some aspects of them, keep changing over time. Gazelles get faster and faster to get away from cheetahs, and cheetahs get faster and faster in response. The genes involved in immune systems are among the most rapidly changing over evolutionary time (Hughes, 1997). But there can be relatively stable adaptations to moving targets as well. For example, many biologists think that sexual reproduction evolved as a strategy to thwart parasites by endlessly recombining genes, creating a moving target (Hamilton et al., 1990; Maynard Smith, 1978b; Tooby, 1982). And immune systems, of course, evolved a combination of fixed and variable features for the same reason: They generate a constantly varying array of antibodies, but the machinery that makes them has many features that have either been conserved over evolutionary time or, in some cases, convergently evolved in different taxa (Ausubel, 2005). You can probably already see an analogy to the types of information-processing systems we've been discussing: systems with constant features that have been designed, by selection, to produce variable elements or to handle open content.

There is now a small but growing body of literature looking at how information-processing systems can evolve in variable environments, and how natural selection can shape them to deal with those environments. But there are some cases in which it has been argued that environments are so variable, and natural selection so slow, that adaptations to those environments are impossible or unlikely—at least via genetic change. Culture-gene coevolution theorists Robert Boyd and Peter Richerson have argued that this is, in essence, why human capacities for culture and cultural transmission evolved: Because culture can change more rapidly than genes, it gives us an additional channel of adaptation that allows us to track changing environments, including social ones, much faster than genetic evolution can (Richerson & Boyd, 2005; see also Tooby & DeVore, 1987, on the notion of "ontogenetic ambushes").

One form of this argument holds that evolved adaptations, because they are innate and frozen in the past, can't keep up with rapid environmental change, so culture stepped in. The frozen-in-the-past bit is probably not right and differs subtly from the idea that culture adds an additional channel of information transmission

to existing genetic ones. In case the difference isn't clear, consider something like theory of mind or our object mechanics systems—classic examples of evolved psychological adaptations. Both probably work fine in the arctic or at the equator. So do food learning systems, kin recognition systems, and the like. Indeed, I can think of few examples of proposed cognitive adaptations that would be rendered more or less effective, in general, at different locations on the globe or in different global climate regimes. Moreover, most of the systems we've considered, like theory of mind and food learning, can handle culturally shaped inputs (thoughts about movies, preferences for chocolate bars). The appearance of culture might, and in many cases probably did, further shape these mechanisms. But as we've seen before, the evolution of new adaptations doesn't necessarily override or erase the operation of older ones. Instead, new cognitive abilities usually add to older ones in a synergistic way. Thus, culture doesn't (mostly) override evolved cognition; it works with it.

As the first law would lead us to expect, the details matter a lot in cases of adaptation to moving targets, whether they are climates, diseases, mates, or cultures. In what kinds of circumstances are changes in the world too fast for evolution to track? In any given case, what are the dimensions along which such change occurs, and are there any dimensions along which there is a stable, trackable signal? And given the notion of open reaction norms we have developed, what kinds of norms might we expect to evolve to deal with moving targets?

Over the next few chapters, we will examine these questions with an eye toward understanding culture as the product of just such a set of mechanisms: adaptations for transmitting information that is highly variable and rapidly changing. The curious and perhaps seemingly paradoxical part, of course, is that in order to work, these mechanisms must make inductive bets about their proper domain. If the proper domain—culture—is virtually defined by its variability, how are any adaptations to it possible at all?

---

We'll begin with some of the conceptual tools that are necessary to understand the evolution of adaptations in changing worlds: evolutionary dynamics and evolutionary game theory.

In chapter 2, we considered the metaphor, first introduced by Sewall Wright, of evolutionary change as hill-climbing. Conceptually, this is a useful metaphor because it points to the fact that some parts of possibility space (gene space, phenotype space) are fitness-better than others, and processes of reproduction and differential mortality—natural selection—tend to expand the uphill sides of population clouds and evaporate the downhill sides, gradually leading them uphill. But as I pointed out, the

idea of static fitness landscapes, while relatively easy for us to picture in our minds, is probably not correct in most cases. The real picture is often more like a landscape in constant upheaval. The reason is that there are many ways for evolutionary changes in a population to feed back on themselves. For example, part of the environment of predators is prey and vice-versa, and an evolutionary change in the design of one changes the fitness landscape of the other. The same goes for hosts and pathogens. And *within* species, changes in design can cause evolutionary feedback as well, because in social species we are an important part of each others' environments. For example, changes in the mating psychology of females can change the fitness landscape for males and vice-versa, leading to evolutionary cascades (Kokko et al., 2002).

In the language used by physicists, interactions within and between evolving populations and their environments are examples of dynamical systems. In physics, the study of dynamical systems originated in the attempt to predict the behavior of systems of mutually influencing objects, in particular, planets. In biology, it's possible to capture the behavior of at least some evolutionary dynamics, including adaptation to moving targets, using mathematical formalisms. As physicists discovered in the case of planets, even some fairly simple sets of dynamical equations can result in highly unpredictable and chaotic behavior. However, the tools of evolutionary dynamics allow us to actually look at the details of "moving targets" and ask when adaptation to them can occur, when it can't, and why.

To get a sense of the simplest possible case of evolutionary dynamics, suppose the frequency of a type in a population changes as a function of two things: its fitness relative to other types in the population and its current frequency. This is a common evolutionary scenario, because the fitness of type A relative to B determines how much better A is at reproducing, but the reproductive rate also depends on how many As there are (fewer As, fewer A babies). Thus the frequency of type A at time 2 equals the frequency of type A at time 1 multiplied by its fitness relative to the population mean (the better you're doing relative to everyone else, the faster your frequency in the population increases). Time 2 then becomes the new time 1, and the process is looped recursively to yield the dynamics of the system over time. This recursive process yields what is sometimes known as the "replicator dynamics" (Taylor & Jonker, 1978).[1] We needn't worry about the precise details here except to note that the dynamics of even such a simple system are not obvious without taking into account evolutionary feedback. A's frequency changes over time, changing the

[1] A general expression for the replicator dynamics described here is $dp/dt = p(W_A - W_{mean})$, where $p$ = frequency of type A, $W_A$ = fitness of type A, $W_{mean}$ = mean population fitness, and $t$ = time. Thus, the rate of change in A's frequency is a function of its current frequency and its reproductive success relative to the population mean.

rate at which A individuals are being produced, which changes the frequency of A yet again. If A starts from low frequency and has higher fitness than B, then its frequency increases in what is known as logistic growth, or an S-curve. Growth starts slow at low frequencies (because there are fewer As to reproduce), accelerates as the number of As increases, and tapers to a maximum level once nearly everyone in the population is A, and A no longer has a competitive advantage.

In this case, then, there is a feedback effect where current rate of change affects future rate of change. But here, fitness is constant. This could occur if, for example, the environment is static. In the evolution of phototactic bacteria, for example, one type, A, might have higher fitness than another type, B, because A is better able to avoid environmental UV light, which is a constant. But what if fitness *itself* is not constant? What if, for example, as type A becomes more common, it alters the environment in some way that changes its fitness relative to other types? Or what if the fitness of A is influenced in some way by its own frequency? For example, in the dynamic known as negative frequency-dependent selection, or rare-type advantage, the fitness of A might be negatively correlated with its frequency, such that A thrives when rare but its fitness decreases as it becomes more common. Thus, it can never completely take over the population. This can occur, for example, if parasites adapt to the most common host genotype. New genotypes will be able to invade the population when rare, leading to continuous evolutionary change.

Here we enter the realm of evolutionary game theory. Game theory was originally developed by mathematicians to model situations in which your best move depends on what others are doing—hence the analogy to games, such as poker or chess, where what strategy is best depends on the strategies others are playing. Game theory was imported into evolutionary biology by John Maynard Smith, who interpreted the idea of "best move" in fitness terms and used the tools of evolutionary dynamics to study how populations of strategies would change in frequency over time when they played each other in populations (Maynard Smith, 1982). Crucially, he introduced the idea of an evolutionarily stable strategy, or ESS. This is a strategy or mix of strategies that can't be beat by any other mix of available strategies. In dynamical terms, it is a stable equilibrium that always returns to that equilibrium when perturbed by, for example, adding new strategies.[2] In many cases, what we might think of as stable adaptations, ones that remain constant over evolutionary time, are ESSs. However, it's important to realize that not all evolutionary strategy spaces have stable equilibria. There are well-known cases where the frequencies of types can fluctuate

[2] Note that any ESS is only an equilibrium against some defined set of strategies; it's possible that an ESS could be defeated by some new strategy the theorist hasn't considered (in the case of modeling) or that hasn't appeared as an actual evolutionary variant (in the case of reality).

eternally, as in predator-prey dynamics (described by the so-called Lotka-Volterra equations), or where they can be even more complexly chaotic, never settling into a constant pattern at all.

We won't delve into the intricacies of evolutionary game theory or evolutionary dynamics (for overviews, see Gintis, 2009; Maynard Smith, 1982; McElreath & Boyd, 2007; Nowak, 2006). Instead, I want to focus on what they add to the language of shapes and spaces I have been developing throughout the book. The first thing to note is that any discussion of moving targets will ultimately need to be addressed using these conceptual tools. They allow us to be more precise—as precise as we like, in fact—about what is changing over time, and how, and why. And indeed, these are just the tools used by theorists of culture-gene coevolution to study how, when, and why adaptation to changing environments occurs (Boyd & Richerson, 1985; Henrich & McElreath, 2003). However, these tools have not yet been fully applied to the study of cognitive mechanisms. This is evidenced by the still-widespread impression that domain-general mechanisms are for variable environments and domain-specific mechanisms are for static ones. I hope to convince you that this is a blunt dichotomy indeed, and that we can do better.

It's difficult to appreciate just how common evolutionary feedback situations are, and how important they are likely to be in the evolution of the human mind. Game theorists sometimes make a distinction between "games against nature" and games against other people. If by "nature" we mean everything nonhuman (or in the case of other species' evolution, everything non-them), then in terms of dynamics, games against nature come in several types. The simplest are games against static environments, such as in the evolution of phototaxis, where there are dynamics in the sense of changes in populations over time, but the adaptive landscape is (mostly) not moving, at least in the dimension of the properties of UV light. Then there are games against changing environments, but ones that are not facultatively responding to changes in the evolving population; for example, polar bears may have to adjust to global warming, but global warming is (mostly) not being influenced by changes in the population of polar bears (we, of course, are a different story). Finally, there are games against nature that are truly game-theoretic in the sense that nature is responding to changes in the population: Predator-prey games and host-pathogen games are examples we've seen.[3]

[3] Technically, games with only one decision-maker might be considered outside the domain of game theory, but still part of decision theory more broadly. If we include other agents as part of nature (e.g., parasites and predators), then some games against nature do have game-theoretic dynamics.

Much of our cognition, including much of human cultural activity, involves games against other people. Not all culture, of course, need involve games against other people. The design of axes is culturally transmitted, but the game we play with axes is against trees. However, many if not most cultural phenomena involve interaction between minds. And here is where we absolutely, positively need to think about evolutionary feedback using the tools of evolutionary game theory. It makes no sense to think of social adaptations of any kind—from theory of mind, to cooperation, to language, to culture—without thinking about the nature of the social feedback environments in which they evolved. When building theories about the design of mental adaptations in the domain of social games, this is going to require some new ways of thinking, because the adaptations themselves, if there are any, *must* be adaptations to moving targets—at least in some cases. So the old way of thinking about mental adaptations as innate and rigid reflexes won't do. Instead, we'll have to develop new formalisms—open reaction norms, adaptations for culture, specialization for moving targets—and take seriously the idea that flexibility and design are friends, not enemies.

---

To begin the discussion of adaptation to moving targets, I'm going to briefly stroll through the minefield known as the evolution of language. If you're like many people, you might have had strong opinions about my mention of Noam Chomsky in chapter 1 (a safe inductive bet if ever there was one). Chomsky has been at the same time one of the most influential and most polarizing figures in the study of language. This is arguably because of the boldness, and some would say absurdity, of his ideas: in particular, the ideas of universal grammar (UG) and a language acquisition device (LAD). There is lots of literature and argument about these ideas, which I mostly have no intention of touching upon here. In fact, I'd like to ask you to forget, for a moment, whatever you might think or know about UG and LAD, and think with me from scratch. What *would* or *could* these things be, biologically, if they did exist?

Remember from chapter 1 that the existence of a LAD was predicated on the idea of a learning problem, and in particular, a problem of opacity; that what is to be learned—including grammar, but also things like the meanings of words, norms for word usage, the meanings of idiomatic expression, and much more—is only manifested in specific tokens (utterances, speech acts). As many before Chomsky pointed out, languages are generative systems wherein a finite number of elements, "words and rules" in Steven Pinker's (2011) concise formulation, can be combined in potentially infinite ways. Chomsky liked to capture this with a phrase attributed to Wilhelm von Humboldt: "infinite use of finite means" (Chomsky, 1972). Now,

we can argue, and many do, about whether native speakers of a language really possess or represent anything like rules of their language (see, e.g., McClelland & Patterson, 2002). For the sake of argument, let's adopt a fuzzy, probabilistic model of cognition and assume that they do have some rule-like generative structures in there. English speakers, for example, have some set of computational procedures for generating utterances in the past tense, like "he tripped and spilled his drink" (these must be variable across English speakers to some degree, though there must also be non-zero overlap if successful communication occurs). The opacity problem, in this case, arises from the fact that children (usually) only hear tokens of the past tense being formed (e.g., tripped, spilled). They usually are never told the rule itself (something like "add -ed" in this case) but must infer it. And while this example might seem simple, when one considers the vast number of principles an adult language speaker knows (embodies) about her language, the learning problem is not trivial.

As I mentioned, the ideas of LAD and UG have taken on many byzantine technical dimensions in the literature, but let's disregard those and take a broad view of LAD as, in essence, *whatever* a child needs to learn language. On this view, the existence of LAD is an empirical fact, because children *do* learn language. From this rather uninteresting definition of LAD, though, there is an interesting question we can ask: Is there anything in a LAD *specific to language*, and if so, what is it?

This question might sound a little crazy, so let me unpack it. If LAD means language acquisition device, and we've defined it as "whatever allows children to learn a language," how could LAD be anything other than specific to language? Here's how: Lots of language learning presumably occurs via mechanisms that did not originally evolve for language learning. Indeed, this is the predominant position among skeptics of Chomsky, who ask if we really need to posit anything *specific* to language to explain language learning (Everett, 2012; Lee et al., 2009). Maybe it can all be done with other-purpose learning mechanisms (I'll assume they have *some* purpose, just not learning language per se). For example, theory of mind could be immensely useful for learning lots of things, *including* language—"what is he *intending* to communicate to me?"—but perhaps it evolved prior to language or without any specific selection for language learning (Bloom, 2000). And other kinds of mechanisms, like statistical learning and neural networks, have been proposed to be able to capture regularities in linguistic input, such as those produced by grammatical rules, without ever having been selected to learn grammar per se (Elman et al., 1996; Kirkham et al., 2002; Saffran et al., 1996; Saffran, 2003). So if we define LAD as "whatever kids need to learn language," it's a cogent position that LAD might contain no language-specific mechanisms, and therefore reasonable to ask what part, if any, is language-specific.

It's worth pointing out, though, that there are several possibilities for what we might mean by "language-specific" here. The most obvious possibility, of course, is that there are some mechanisms in LAD whose proper domain is language learning and that alone, and that selection built them for that purpose. Another possibility, though, is that language learning exerted selection on some learning mechanisms that *do* help in learning language, and have been selected to do so, but also help in learning other things, and might have been selected to do that too; theory of mind mechanisms are one such possibility that we'll discuss more below. A final possibility is that there might be some mechanisms—some aspects of mental structure—that make language learning easier, but not because they have been selected to do so. One way this could happen is if languages evolve through a process of cultural evolution to become more learnable to the mind. On this view, the informational nooks and crannies of the mind predate language, and languages that are successful at being learned by a mind with those nooks and crannies are the ones that stick around. If true, a mind with *different* nooks and crannies wouldn't be as good at learning the human languages that actually exist, and it certainly might look that the nooks and crannies had been designed to acquire the languages that are there.

Just such a proposal has been made by language scholars Morten Christiansen and Nick Chater (2008). They argue that at least one aspect of languages—their grammars—constitute an evolutionary moving target. This is because grammars, being transmitted culturally, change faster over evolutionary time than genes. Thus, they argue, any kind of genetic adaptation specifically for acquiring grammars (i.e., a UG) is impossible. Instead, they argue, languages evolve to fit the mind, and so we should explain the apparent ease with which most children acquire their local grammar not by appealing to any language acquisition adaptations, but rather to a process of cultural evolution in which features of languages that make them more learnable are selectively retained over many generations of transmission.

Christiansen and Chater's argument is a technical one backed up by evolutionary modeling, and my goal is not to delve into it in detail but rather to use it as a conceptual springboard for thinking about adaptation to moving targets. Certainly, the idea that languages evolve to fit the mind makes sense. Indeed, this is likely to be a general feature of the products of cultural transmission, because they must pass every generation through the sieve of human minds. But is it true, as Christiansen and Chater claim, that if B evolves faster than A, then A can't adapt to B? And if so, what does this imply for the study of mental adaptations?

First, let's grant the argument that cultural change is, in fact, faster than evolutionary change in genes or gene regulation systems. A recent paper by anthropologist Charles Perreault comparing rates of change in human-made artifacts with rates of

change in the morphological traits of animals lends support to this idea (Perreault, 2012). As a general principle, it could be true (though probably not without exceptions). And, if Boyd and Richerson are right, it's the whole raison d'être of culture. But a glance at other areas of biology suggests that it can't be a blanket rule that if B evolves faster than A, then A can't adapt to B. Otherwise we'd be extinct many times over, because pathogens are trying to kill us all the time, and they can evolve orders of magnitude faster than we can. We clearly do have adaptations to thwart pathogens—not perfectly, of course, but enough to keep up our end of the fight—despite potentially massive differences in rates of evolutionary change. These adaptations include, of course, the immune system, a combinatorial system designed to generate a broad range of antibodies, with inductive bets about the range of variation in pathogen proteins (Kauffman, 1993). This is not to say that pathogen resistance and language learning are the same problem, of course: The immune system has had a far longer time to evolve, it involves interspecies coevolution, and antigen detection and language learning are different in many other ways. However, this is another case where the details matter: It matters what's changing, and what's adapting to it, and why.

Christiansen and Chater's argument has specifically to do with grammar. They are not arguing that there can't be some adaptations for language acquisition, though they suggest (and many others seem to agree) that many mechanisms involved in language acquisition probably predate the evolution of language. What they are arguing with is the idea that grammatical rules, or the space of possible grammatical rules, could be genetically encoded.

They might well be right; intuitively, the idea of grammatical rules being genetically encoded, whatever that might mean, seems unlikely (we've already discussed at length problems with the idea of pre-specification). But let's look at the general grounds for the claim that there can't be genes that help us acquire grammar—or anything else that presents a moving target, for that matter. They claim that their argument is an a priori one and that a UG is, in essence, logically impossible. They base this on a model where languages culturally evolve with a population of learners learning and transmitting them (Chater et al., 2009). Genetic mutations are occurring, and genes that make existing languages easier to acquire give a small fitness boost and are selected for. This is a model of what is sometimes called the Baldwin effect (Simpson, 1953; Weber & Depew, 2003). In the Baldwin effect, an aspect of the phenotype, such as language, is first acquired through some general-purpose process, such as learning—there has been no selection specifically to acquire it. Then, if genes appear that make acquiring it easier, they can be selected for, leading to adaptations specifically for acquisition (or, under some interpretations, innateness). This has widely been proposed as a scenario for language evolution (e.g., Pinker & Bloom, 1990). However, Chater and Christiansen's models suggest that, under a fairly wide range of assumptions, the Baldwin effect has trouble catching up

with cultural change in languages over time.[4] Thus, they argue, cultural adaptation of languages to make them more learnable, given the brain's preexisting mechanisms' constraints, is more plausible than the evolution of genes specifically for acquiring language that already exist.

There are several things to note here. One is that their model assumes that languages, in the beginning, are already learnable—they are present, with some degree of complexity, in the model's first generation. Thus, the moving target preexists, and genes are under selection to catch up with it. But is this the only way language evolution could occur? For one thing, it seems to assume that the major problems that evolution needs to solve—such as coordination problems, to which we'll return in a moment—have already been solved by the time natural selection enters the picture. This is a potentially problematic assumption. The model also assumes that the major source of selection on language genes has to do solely with improving the learning of languages that already exist; the fitness function is solely based on speed of acquisition. But it's also at least as possible, if not more so, that genetic changes related to language have been favored because of the expressive possibilities they *enable*—something not captured in a model that conceptualizes the complexity of language as predating any genes selected to produce it. Selection isn't acting on genes just because of their effects on learning, but also because of the fitness benefits that come from the expressive possibilities of the languages they enable the child to produce. Language learning and language production are therefore two sides of the same coin, or phenotype, and it's not possible to place one before the other.

To assume that linguistic complexity exists and then genes evolve to acquire it is, of course, one possibility, but it's also possible to imagine linguistic complexity itself as evolving via a stepwise process in which small variants in brain design, due to changes in genes or gene regulation, enable small nudges up the linguistic complexity hill, both in production and acquisition (Jackendoff, 2002). While it's an open question, empirically, there is no reason to assume that grammar exists *before* genes for it. Grammatical innovations, such as the use of word orders to disambiguate meaning (e.g., man bites dog vs. dog bites man), could be enabled by small changes in brain design that make them possible. Not directed changes, of course. But brains are constantly varying a little in design space, and fortuitous changes that allow richer varieties of meaning to be communicated could be selected for.

Biologist Martin Nowak and colleagues have modeled the evolution of syntax by asking when the benefits of syntactic combination of words—nouns and verbs, for

[4] Note that they make this argument for "arbitrary" as opposed to "functional" features of language. Many grammatical properties of languages, like word order, are sometimes held to be arbitrary. We'll return to this issue below.

example—exceed the benefits of simple symbol-to-event mappings (non-syntactic communication) due to the expressive benefits of combinatorics (Nowak et al., 2000). They found a tipping point of linguistic complexity above which syntactic combination is favored. While they only modeled the simplest possible kind of syntax—dyadic combinations of two symbols—their approach can be applied to more complex syntaxes evolving cumulatively in a stepwise fashion. This does not mean, importantly, that all the changes in syntactic complexity need be *genetic* changes—but equally importantly in their model, there is no reason why accumulating syntactic complexity wouldn't proceed via gene-culture coevolution rather than via genes locked in place and culture adapting to them.[5]

One problem in thinking about language evolution is what appears to be a "first mover" or "chicken and egg" problem. Let's say some mutation allows me to use word order to disambiguate meaning, so I can mean something different by "dog eat" and "eat dog." If I'm the first person in whom that mutation occurs, how can it ever be selected for, since nobody else will understand me or be able to learn my innovation? As it turns out, this is a problem that bedevils the evolution of any adaptation where, metaphorically speaking, it takes two to tango. This is best known in game-theoretic work on the evolution of cooperation. It has been known since William Hamilton's work on kin selection in the 1960s that altruistic types can spread if they help copies of the same altruistic type in the population. But when the mutation for the type first appears, who is there to help? This is a difficult and interesting problem, but models of the evolution of cooperation are beginning to show that it is not an insurmountable one, because early in the evolution of the new innovation, multiple copies can encounter each other by chance—essentially, through the genetic lottery of drift, before the gene has begun to be selected for—creating the spark that allows the innovation to take off. Empirically, the problem must be surmountable because cooperation has, at least in humans, evolved—as has language, a special form of cooperative behavior.

An interesting feature of language games is that they are, at least in part, *coordination* games. These are games in which one party's behavior affects the other

[5] Chater et al. (2009) do model a case in which language change is determined by both genetic and cultural evolution. However, the model is akin to a walk through arbitrarily varying grammars of equal complexity, and fitness is determined solely by how long it takes to learn the grammars. One can imagine instead, as modeled by Nowak et al. (2000), a model of the evolution of increasingly complex grammars in which fitness is also determined by the expressiveness of the language one is learning—which is, presumably, the primary function of the ability to combine symbols syntactically. If accumulating linguistic complexity occurs in part because of fitness benefits of increasing expressivity, then a crucial question for the evolution of grammar adaptations is whether any genetic changes are selected for because of the boost in expressivity they enable, not just their effects on learnability.

and vice-versa, and in which both parties can improve their fitness by choosing a particular behavior based on what the other will do. A well-known real-life example is driving on the right or left side of the road. As an individual, it doesn't really matter whether you drive on the right or left; it's arbitrary, and the choice of one versus the other is seemingly neutral, or nonfunctional. But as soon as there are other drivers, it matters a lot which side you drive on: It should be the same side as everyone else. The solution, in such situations, is sometimes called a coordination device or, more commonly, a convention: an arbitrary agreement to do it one way instead of the other, even though neither way is objectively better and could be equally well done differently someplace else (Carnap, 1937; Clark, 1996; Lewis, 1969; Tomasello, 2003). Languages contain many conventions, from the rules of grammar to the meanings of individual words. For example, it doesn't really matter which pattern of sounds a language uses to mean "cow," but speakers of the same language need to converge on the *same* pattern if they are to be able to talk about cows.[6]

What's interesting about languages, and cultural phenomena more generally, is how replete they are with conventions that are both arbitrary and functional (Pinker & Bloom, 1990). A convention like driving on the right side of the road is arbitrary in that the left would also work fine. However, it's also most definitely functional, in the sense that picking one of the two possible conventions saves lives. Similarly, every language has many conventions that could easily be different—and are, in different languages, as in the world's many words for cows. And yet these conventions are functional in that without them, linguistic communication wouldn't work.

An important question for us is whether there might be adaptations that assume conventions exist, and try to figure out what they are. In the case of language, for example, learning systems could make the inductive bet that word order matters—thus paying attention to it—and trying to figure out the word order for their particular language (this inductive bet would sometimes fail, since not all languages use word order to disambiguate meaning, though it is a common convention; see Evans & Levinson, 2009). Another example that we will visit below is the assumption that one's culture has social norms that must be learned in order to be a proper citizen. These might be arbitrary and yet critically important to fitness. As yet, we don't really know if there are learning mechanisms specialized to do this. But in the case of language, we might expect learning mechanisms that are open along the many

[6] There may be, as usual, some exceptions. In the phenomenon known as "sound symbolism," the sounds of words bear some non-arbitrary relationship to their meanings (Ohala et al., 1994). For example, words for little things might sound little, or words for sharpness might sound sharp. Not all conventions need to be entirely arbitrary; it's just that they can be.

dimensions in which languages vary, and yet that still make inductive bets about their proper domains.

---

The question of whether there are genes selected for language acquisition is still an open one—not to mention the question of what *aspects* of language there might be adaptations for. The larger point is that moving targets present an interesting, but not necessarily insurmountable, problem for evolution. Whether adaptations to moving targets can evolve depends on the details.

There is an interesting feature of language evolution and the evolution of culture more generally that makes it quite different from other cases of coevolution that we've seen, such as predator-prey and host-pathogen coevolution. Host-pathogen coevolution is antagonistic: it is an evolutionary arms race in which each species is evolving attacks and defenses to the other. In culture-gene coevolution, of which language evolution is a special case, the coevolutionary process occurs within a single species and likely to be mostly cooperative or synergistic rather than antagonistic. Whereas host-pathogen coevolution is an arms race where each party is trying to run away from the other's adaptations, languages and genes that might help acquire and use them are not, in any metaphorical sense, trying to run away from each other. On the contrary, they are likely to be running toward each other: Genes are selected to mesh synergistically with languages and vice-versa. Whatever languages exist must already have been learnable, even when they first appeared, and selection is only likely to improve that situation, whether on the gene side of the interaction or the language side.

Interestingly, it is possible to have within-species antagonistic coevolution. Certain aspects of interactions between males and females, for example, may entail antagonistic coevolution when fitness interests diverge (Rice, 1996). And in humans, it has been proposed that we have been selected to become progressively more intelligent in order to outsmart each other, an idea sometimes known as the Machiavellian intelligence hypothesis (Byrne & Whiten, 1988). Among other things, there might have been selection to try to figure out others' strategic goals, plans, and schemes—possibly a major source of selection for theory of mind. Here again, we have a moving target, and the dynamics probably matter for the kinds of adaptations that might evolve as a result.

An interesting feature of these kinds of coevolutionary scenarios is that they can be autocatalytic, or self-feeding, leading to constant change. Some dynamical systems are self-stabilizing, settling into equilibrium. Others, like arms races, can have "runaway" dynamics that lead to an ever-increasing spiral (Kokko et al., 2002). If the

Machiavellian intelligence hypothesis is right, then maybe we will keep getting ever smarter because of the incremental fitness advantages of being the smartest person in the room.[7] Language evolution is interesting because one might expect stabilizing dynamics that make languages more learnable. If, however, the possibility space for adaptive linguistic complexity is large, then we could still be on the side of a fitness hill, with further innovations in the possible languages human minds can create yet to come.

And then there is cultural change. Empirically, the history of cultural evolution certainly looks like a kind of runaway process, starting slowly with the stone tools of several million years ago and then picking up steam in an ever-accelerating process of cultural and technological change that is now unfolding at a frenzied pace. If ever there was a moving target, this seems to be one. And yet it is clearly a moving target created by ourselves and our minds; culture and minds can't be decoupled, or one understood without reference to the other. However, the relationship between minds and culture remains enormously contentious. Many anthropologists and psychologists treat culture as an independent causal force on development, and many "nativist" psychologists treat culture as something that is mostly poured on top of innate mechanisms without altering them. If culture and the mental mechanisms that enable and transmit it coevolve and are therefore deeply causally intertwined, what are the implications for understanding mind design?

[7] Or perhaps not. One could imagine technological crutches, like computers, relaxing selection on some aspects of intelligence, or at least altering the dimensions of intelligence that are being selected for. This debate is already playing out in parents' concerns about the effects of technology on their children.

# 9

# CULTURE

Many organisms, including virtually all animals, transmit information socially. A dog urinating on a fire hydrant is transmitting information socially, as is a bird that takes off from a telephone wire and scares other birds into doing so. Monkeys warn other monkeys about predators via special alarm calls (Cheney & Seyfarth, 2007); rats acquire food preferences from other rats by smelling their breath (Galef, 1996); and female guppies observe which males other females have mated with and copy their choices (Dugatkin, 1992). And in some cases, there are even things that begin to look like genuine culture, as when chimpanzees learn to use sticks to fish for termites, or capuchin monkeys use odd, social greeting rituals like poking their fingers into each others' eyes, something vaguely and more painfully like the monkey equivalent of a handshake (Perry, 2011; Perry & Manson, 2008).

But human culture is categorically different. While other animals can scare each other, make friends, gang up in teams, and even sometimes create culture-like products such as songs, rituals, and tools, only humans have invented calculus, built cathedrals, written *Moby Dick*, and landed on the moon. The cultural products we create are incredibly complex, a huge amount of our time is devoted to processing and interacting with them, and they structure virtually every waking moment of our lives. Why?

One reason is that human culture, unlike the culture of other animals, is cumulative. Even the most sophisticated cultural products of other animals never get beyond a certain complexity. Although there are some very clever monkeys and chimps who might invent new ways to hold a termite-fishing stick or remove the spines from a seed, chimps do not, to use Isaac Newton's phrase, see farther by standing on the shoulders of giants. The products of animal culture do not ratchet ever upward in complexity resulting in exponential technological change (Boyd & Richerson, 1996; Tennie et al., 2009).

Human culture, on the other hand, accretes. It builds on itself and climbs toward ever-higher levels of complexity: It evolves. And, as for every other adaptive property we've examined, it turns out that this is not a property that comes for free. It must be enabled. Cumulative cultural evolution is only possible because

of specialized adaptations that evolved *because* they produced culture that could accrete, where new innovations could build on older ones (Boyd & Richerson, 1996). And this is only possible because the evolved mechanisms that enable cultural transmission have design features, fit characteristics that enable them to do what they do.

What is perhaps odd or counterintuitive in thinking about these mechanisms is that their EEA, the environment to which they are adapted, is culture (Boyd & Richerson, 1985). That is, they are adapted to acquire and transmit information that is by nature open in many, many dimensions and constantly changing. By now, I hope, this is not so strange to think about. Indeed, we have already seen cases of mechanisms designed to transmit and acquire cultural information, namely, the mechanisms underlying mindreading and, possibly, language. Cultural transmission mechanisms are also responsible for our abilities to use tools, cook food, make clothing and shelter, read, do math, and produce television shows. They are paradigm cases of mechanisms whose inputs and products are, and have always been, evolutionarily novel tokens. But, in order to have evolved, these mechanisms must have been selected to embody assumptions about *types* of information. In other words, there must be design features that account for the fact that humans have cumulative cultural evolution but monkeys do not. What could they be, and what kinds of ontological commitments could they possibly have to their proper domains that allow them to do their job?

Here, perhaps, the imagination runs up against a brick wall. What could igloos, the Taj Mahal, calculus, and *Seinfeld* possibly have in common? Isn't culture the archetypal moving target, the very embodiment of unpredictability and novelty? As it turns out, there are formal models of cultural evolution that look at the properties of transmission mechanisms that are necessary for cumulative cultural evolution to occur and that examine the fit between transmission mechanisms and environmental circumstances that make them adaptive. Interestingly, just as was the case with language—a special case of culture in some ways, an ordinary case in others—cultural transmission mechanisms in general can be seen as adaptations to their own products. The acquisition and production mechanisms in one individual are adapted to acquire and produce what is being created by those same mechanisms in other individuals. And once such mechanisms are in place, they can have radical effects on the evolution of the organisms that possess them. In the case of humans, we have become downright culture-dependent. Our minds house a host of mechanisms that *assume* cultural information will be out there in the world, from language to human thoughts to the artifacts we create, and actively attempt to absorb it.

Beginning with the early models of the evolution of cultural transmission by Robert Boyd, Peter Richerson, and others in the 1980s, the study of the formal properties of mechanisms of cultural transmission and the conditions under which they can evolve has become a burgeoning field (Boyd & Richerson, 1985, 1988, 1992, 1996; Feldman et al., 1996; Franz & Matthews, 2010; Giraldeau et al., 2002; Henrich & Boyd, 1998, 2001; Henrich & McElreath, 2003; Laland, 2004; Mesoudi, 2011; Mesoudi & O'Brien, 2008; Rendell et al., 2010). In general, these models use the logic of population genetics, evolutionary game theory, and epidemiology to examine the conditions necessary for cumulative culture to evolve, and the evolutionary dynamics that occur once it does. What they model is the evolution of what you might think of as transmission rules. These rules are instantiated in psychological mechanisms; what evolves are developmental systems that build these rules into peoples' brains. And as is the case for any psychological mechanism to evolve, this requires input conditions, rules or principles of processing, and outputs that other systems—the systems designed to shape our behavior based on cultural information—can use.

Consistent with the view I was advocating in the last chapter about language mechanisms, cultural transmission mechanisms simultaneously enable the acquisition of culture and the production of it. In formal models, this means that an individual who acquires a culturally transmitted behavior can then produce that behavior. One can also think of individuals as acquiring knowledge and then generating behavior based on that knowledge; this would characterize, for example, learning the rules of language and then producing utterances, or learning how to make a fishing pole and then making one. Culture acquisition mechanisms evolve because their products—culturally transmitted behaviors—impact fitness. Culture, in this case, is just a word for the knowledge and behaviors these mechanisms acquire and transmit, and can be a broader class of things than what is often implied by the everyday notion of culture. For example, the way you tie your shoes and the way you blow your nose are part of culture.

Any kind of learning implies frame problems. Of all the things we *could* learn from the soup of information around us, we only *actually* learn what our cognitive mechanisms make learnable. Cultural transmission mechanisms evolve because they enable certain kinds of learning, and this in turn means that they solve frame problems. They cause what is to be learned to pop out, to become apparent and learnable to the organisms that possess them. Pondering what this means, and what design features might allow cultural transmission mechanisms to get this done, is the key to understanding how they are fit to the world of cultural information.

I've mentioned several examples of how learning mechanisms cause this popping out to organisms that have them. For an indigo bunting that has gazed long

enough at the sky, his route home pops out at him—it is painted onto the stars. The same doesn't happen for us no matter how long we gaze (except in the case of certain culturally transmitted traditions of celestial navigation). Similarly, for a human baby who grows up in a home where language is spoken, the meanings that others are trying to convey eventually pop out for her and continue to do so for the rest of her life. People speak and meaning is delivered. This does not occur for a cat growing up in the same home (the cat can learn some sound-meaning associations, of course, but never parses sentences or comes to appreciate poetry).

Consider, now, the difference between a capuchin monkey observing another monkey breaking nuts with a stone and a person observing the same event. What do the two parties see? At some level, the human and the monkey see the same thing: the same objects, the same patterns of motion. But there is something very different about the learning mechanisms in the two species, because the monkey takes a very long time to learn anything even approximating the behavior, going through many random variations of grasping rocks and nuts and doing things with them (Ottoni et al., 2005). A human child, on the other hand—while not getting it perfectly, of course—begins to imitate much more closely even on the first try, seeming to have grasped something about what the person is *trying* to do. The child detects a relationship between what the person is trying to do—which is, of course, inferred from the action—and the motions they are making. She is then motivated to try to reproduce what she's noticed. This implies a system with at least three features: It attends to what other people are doing, it is able to parse what they are doing in ways that either allow them to imitate it or to acquire some other relevant information about it, and it is motivated to care, motivated to learn something based on this observation. In fact, in order to work properly, the learning mechanism must cause these learning opportunities to be *intrinsically* motivating. The learner can't wait to find out if what she is learning is actually fitness-useful, or she will miss the opportunity to learn. And she will almost never have direct evidence that this is the case at the time of learning, so instead she must rely on other cues, indirect cues, that what she is learning is useful—a classic case of an inductive bet (Boyd & Richerson, 1985). What might such cues be?

Cultural transmission mechanisms are adapted to the environment that they themselves shape: the cultural environment. Boyd and Richerson have shown that this environment has some unique self-structuring dynamics that cultural transmission mechanisms can exploit as inductive bets about the fitness-usefulness of information. This is because, although any given individual might not (and usually doesn't) know the fitness value of any bit of cultural information he acquires, the fact that these bits of information *do*, in fact, have fitness value—some must, or the

system wouldn't evolve—affects the way that information ends up being distributed in a population. Evolved culture acquisition mechanisms can evolve to exploit this fact.

The most obvious fact about how the fitness value of cultural information feeds back on its distribution in a population is that fitness-good cultural variants increase in frequency, and fitness-bad ones decrease. Remember that under the view of phenotypes I gave earlier in the book, one's cultural knowledge is a part of one's phenotype. So the fact that one's fitness can be affected by one's cultural phenotype is entirely analyzable using the Darwinian logic of what happens to variants that have positive or negative impacts on their own rates of reproduction. Of course, for Darwin's syllogism to hold, it must be the case that individuals with fitness-good cultural knowledge or beliefs leave more offspring than those with fitness-bad beliefs, *and* that offspring are likely to resemble parents along dimensions of belief.[1]

Boyd, Richerson, and others have modeled just these conditions and find that under certain circumstances, there is a reliable cue to how fitness-good a particular cultural variant is: its frequency in the cultural environment, compared to other cultural variants (technically, this is true under certain conditions of population structuring; Boyd & Richerson, 1985; Perreault et al., 2012). This is an aspect of environment structure that can select for a particular kind of inductive bet: If I acquire the most common cultural variant in the population, it is likely to be fitness-better than the less common variants. Mechanisms that instantiate this inductive bet produce what Boyd and Richerson call *conformist transmission* or *conformist bias* (also called a frequency-dependent bias; Boyd & Richerson, 1985; Henrich & McElreath, 2003). In essence, these mechanisms are using commonness as a cue, and thereby solve a frame problem: Of all the variants out there that *could* be learned, learn this one, because it is likely to be the most fitness-useful.

Note that if such a mechanism is to evolve, it must have certain other design features as well. For example, it must be able to sample cultural variants over a population of individuals it encounters and keep some running statistics about them, computing which one is more common. There are additional features that this, in turn, implies. For example, the mechanism's certainty that it has found the most common variant should be a function of the number of individuals it has sampled, a basic consequence

[1] Does all culturally transmitted information need to come in the form of beliefs? Not necessarily—and certainly much or most of what we transmit isn't represented in a consciously accessible, propositional format. Even things like styles of walking can be culturally transmitted. I'm using the words "knowledge" and "beliefs" here to stand more generally for any information transmitted culturally, including preferences, dispositions, etc.

of Bayesian sampling theory. This could, in turn, lead to a sample size effect in individuals' willingness to learn something: The more people you see doing something, the more it seems like a good idea to do it yourself (Perreault et al., 2012).

Now, you might be inclined to think about something like a conformist learning mechanism as a paradigm case of a domain-general mechanism. You certainly may, and you'd be right in the sense that the mechanism could operate on a very, very broad class of information. But you'd be mistaken if you thought either that conformist learning mechanisms aren't specialized, or that they didn't have a proper domain, a set of circumstances in which they were designed to operate. In fact, conformist transmission mechanisms require their EEA to be structured in a certain way in order for them to evolve: Frequency must correlate with fitness (Boyd & Richerson, 1985). The mechanisms have input conditions, using cues (local commonness) to admit information for processing. They ignore large amounts of information (the uncommon variants and everything else that has not been parsed as a cultural variant at all). It is an interesting feature of these mechanisms that in a certain sense, they do not care about the content of what is being learned—that is left unspecified. Nevertheless, they are adaptively specialized and exhibit mind-world fit, instantiating inductive bets that are well suited to the environments where they evolved.[2]

---

Frequency is not the only thing that can indicate the fitness-usefulness of a cultural variant. Commonness is a kind of cue to success, but as it turns out, there is another: success itself. In case you think this is circular, consider this: The things that *cause* success are not the same things that *indicate* success (Boyd & Richerson, 1985). This is true all around the biological world. For example, look around you, and you will know that squirrels are successful, in fitness terms, relative to many other animals that could have ended up populating our cities, such as prairie voles or agoutis or owls. But do you know why? Probably not.

The same holds for success within human societies. Some individuals have either directly observable fitness—lots of children—or properties that are probably closely tied to fitness, such as lots of resources to feed those children, lots of family, lots of friends, lots of status and power. Often, perhaps most of the time, we don't know *why* these people have reached their positions (Donald Trump, anyone?). But Boyd

[2] Note that as for other potentially domain-general properties, like the properties of Bayesian learning algorithms, these properties *could* be entirely content-general, or they could also be built into multiple, specialized learning systems in the mind that take cultural inputs. For example, a learning system specialized to learn about danger from conspecifics could use frequency of danger beliefs as a cue for learning about danger (Barrett & Broesch, 2012).

and Richerson have shown that under certain circumstances—in particular, when there is an environmental correlation between the traits that indicate success and the traits that cause success—mechanisms can evolve that cause people to try to imitate the successful, and to acquire their particular cultural variants even if they are not the most common. Boyd and Richerson call this *prestige bias*, a form of model-based bias: You acquire a cultural variant based on the traits of the person demonstrating it, not the commonness of the variant itself (Boyd & Richerson, 1985; Chudek et al., 2012; Henrich & McElreath, 2003).

Interestingly, this form of transmission can lead to a runaway process in which people copy things that aren't in fact useful, leading to a host of non-adaptive outcomes. Or rather, they are not necessarily non-adaptive but are self-feeding, in that the variant becomes linked to success by virtue of the fact that people have prestige-biased copying mechanisms. A good example of this might be the runaway dynamics of clothing trends, where clothing styles can explode into popularity (and eventually extinguish) because of who is wearing them, even though it is not the clothes themselves that necessarily had anything to do with the success of the popularizer. Copying the style *can* have fitness benefits, and these can be arbitrary in nature (e.g., gaining friends or mates because they like your hat). But they need not, and indeed, prestige-biased copying can be bad for fitness in any case where what you copy negatively impacts survival or reproduction (e.g., copying the drug habits of Keith Richards, when you're not Keith Richards).

Such mistakes are an inevitable problem faced by all mechanisms that have to deal with inferential opacity: cases where the link between the underlying property being inferred and the surface cues that indicate it is obscured or opaque (Gergely & Csibra, 2006, 2011). In mindreading, for example, people can make mistakes because of opacity, such as assuming an act was intentional when it wasn't (and potentially paying fitness costs as a result, e.g., getting in a fight because of a misunderstanding). But the opacity problem appears to be especially bad for cultural transmission mechanisms that use frequency or success as cues, particularly if these mechanisms make *no* further assumptions about the link between the underlying cultural variant and its frequency or success. Bizarre, seemingly non-adaptive or maladaptive cultural phenomena such as rain dances or believing in God might well be the result of an inherent opacity problem in cultural transmission (Richerson & Boyd, 2005; Sperber, 1996). The reason is that cultural transmission mechanisms are betting that the underlying correlation structure is the right one (e.g., that the success of people believing something, or the commonness of the belief, are cues to the fitness-goodness of that belief). Often, that bet is wrong—especially because cultural transmission is a process that feeds back on itself dynamically. If people start to copy something because it is illusorily correlated with prestige—it is not the *cause* of prestige, only

accidentally associated with it—it can spread and even become common, causing odd, maladaptive dynamics. Of course, to both evolve and be maintained, the net cost of the maladaptive baggage they accrue must ultimately be less than the fitness benefits they provide. But the occasional jackpot of learning something that has a huge fitness payoff can sustain learning lots and lots of stuff that is useless or even worse.

---

The fitness-goodness of culturally transmitted information can be indexed by statistical properties, such as how common it is in the population or how successful the people who possess it are. But it can also be indexed by what it is *about*. Mechanisms that evolve to exploit the content of culturally transmitted information as a cue to its usefulness are sometimes said to instantiate *content biases* (Boyd & Richerson, 1985; Henrich & McElreath, 2003). Like all learning mechanisms, these are prepared learning mechanisms, with input conditions that cause them to preferentially attend to and retain certain *kinds* of information from the vast ocean of cultural information that bombards us. You can think of them as either filters or sieves that catch what's most useful and throw back the rest, or even better—given that *all* cultural information has some chance of being noticed and retained—as lenses that magnify the importance of some information because of what it's about.

As an example, consider the case of danger. Information about what's dangerous is certainly potentially relevant to fitness. And information about danger is all around you, in culturally transmitted forms: There is your mother telling you, "Billy, don't get too close to the railing," there are news reports about plane crashes and swine flu, there are warning labels on the cleaning products under your sink, and there is the friendly advice about which neighborhoods to avoid when visiting a foreign city. On one hand, the fact that information is *about* danger seems to be good-enough reason to pay attention to it. But on the other hand, given that it's culturally transmitted, you don't know for sure if it's true. In fact, you usually can't *ever* know for sure. You could, for example, spend some time in various areas of a city to see which are more dangerous. But then, of course, you'd face various well-known problems of statistical inference, such as the potential errors of small sample sizes and the difficulty with estimating the probabilities of rare events, such as being attacked.

As it turns out, these are exactly the kinds of adaptive problems one would expect to come into play in the evolution of cultural transmission mechanisms. Boyd and Richerson's models have shown, for example, that it can be fitness-good to attend to culturally transmitted information when two conditions hold: when it's less costly to get the information from somebody else than to find it out yourself—taking the

advice about bad neighborhoods rather than going there and sampling them personally—*and* when the information is reliable or, in statistical terminology, when it has sufficiently high validity to make it predictive above chance, and at the level that yields a fitness payoff (which in turn depends on environments that don't change too rapidly). What makes culturally transmitted information valid, in turn, is the cultural analog of natural selection, which weeds out bad variants—bad or unpredictive information—in favor of good ones.

What this means is that prepared cultural learning mechanisms can evolve when culture is a fitness-better source of information than individual learning: The information type tends to be stored in the minds of conspecifics and indexed by their behavior; it tends to be valid; and the benefit/cost ratio of relying on it outweighs the benefit/cost ratio of individual learning. Under these conditions, learning mechanisms with properties that cause them to systematically attend to certain information types in the barrage of cultural information flow, and to process them in certain ways, can be selected for.

We've already seen one example of this: the acquisition of food preferences by Norway rats, who smell each others' breath to shape their own preferences about what is and isn't good to eat (Galef, 1996). Note that while this information is transmitted socially, it's not the same as a norm. The fitness effects of eating poison come only from an interaction with the world (a game against nature), not an interaction with anyone else. The interaction here is just what shapes the learning. Note also that this is potentially a risky strategy, because it only takes into account what another rat just ate, not the effects on the rat; the rat could have eaten something that will end up killing it in a few hours. And, as I mentioned in chapter 3, it turns out that rats' food updating doesn't take into account whether another rat dies. This strategy depends the role of natural selection in shaping the structure of the social environment: If eating food X tends to make you die and eating food Y doesn't, food Y will tend to be much more common on others' breath than food X, lending statistical validity to the breath cue (Noble et al., 2001). Such social learning mechanisms can evolve only when there is added fitness value above and beyond individual learning, tasting things yourself and seeing if you get sick. Rats do have individual food learning mechanisms, as I mentioned, but presumably the reason they have a specialized mechanism to sample others' food choices is because it gave them a big enough fitness boost above and beyond the risky strategy of tasting things for yourself (Boyd & Richerson, 1985, 1988; Feldman et al., 1996; Rendell et al., 2010).

A similar logic might hold for the acquisition of fear. In a set of experiments, psychologists Susan Mineka and colleagues showed that rhesus macaques acquire fear through social transmission in a manner similar to rats' acquisition of food preferences: They appear to be prepared to attend to certain *types* of information and to

learn it through social cues (Cook & Mineka, 1989; Mineka et al., 1984). Mineka and colleagues studied lab-reared macaques who had lived their entire lives inside a lab and so had very little experience with objects they might encounter in the natural world, such as snakes and flowers. As it turns out, naïve (inexperienced) macaques show no fear of snakes or flowers when they first see them. But Mineka and colleagues showed that these monkeys acquire fear of snakes immediately, in a single trial, when they see another monkey showing a fear reaction toward a snake in a video (in fact, they acquired fear of either a toy snake or a toy crocodile in this manner; Cook & Mineka, 1989). Through a clever bit of experimental manipulation—cutting out the snake and pasting in a flower, but leaving the other monkey's fear reaction the same—they showed that naïve monkeys do not become afraid of flowers or toy rabbits under the same circumstances. This suggests that the monkeys have a prepared social learning mechanism that has at least two input conditions: There must be a fear response by a conspecific, and the object has to be of a certain kind, or it must exhibit certain cues. What those cues are—what distinguishes snakes and crocodiles from bunnies and flowers—is, to my knowledge, still unknown. There is evidence that other animals, such as wallabies, exhibit category specificity in learning about predators (e.g., extending learned fear of foxes to cats but not to goats; Griffin et al., 2001). The nature of the underlying inductive template—what dimensions of similarity or difference it uses—is an important topic for future work.

The design of the social fear learning system in macaques makes sense for two reasons: the extreme potential costs, or risk, of finding out for yourself what is dangerous, and the availability of conspecifics' experience of the world as a means of bypassing at least some of the costs of individual learning. As you might guess, pure information contagion alone is not enough to make this strategy evolutionarily stable; at least some individuals must do the experiment of finding out whether fear of snakes is really fitness-good. But the rats provide an example of how the mere survival of the individuals with the fear might be enough to support above-chance statistical validity. Evolutionary modeling has shown that a mix of individual and social learning is required for such a prepared learning system to evolve, with the nature of the mix adjusted by the relative costs and benefits of the different kinds of learning (Boyd & Richerson, 1985, 1988; Feldman et al., 1996; Rendell et al., 2010).

Based on this logic, I conjectured that humans might possess a similar system (Barrett, 2005a). Indeed, such a prepared social learning system might be widely phylogenetically distributed in primates and still present, though perhaps modified, in humans. In fact, given the wide range of ancestral environments occupied by humans, a prepared cultural learning system for danger might be even *more* important than for other species.

Some species are known to have templates (a set of prepared cues, complexly combined) that allow them to recognize certain predators without any learning at all. I mentioned island-dwelling wallabies that show fear of foxes that have not preyed on them for thousands of years (Blumstein & Daniel, 2005; Griffin et al., 2001). Similarly, psychologist Richard Coss and colleagues have shown that ground squirrels show a fear of snakes, even in snake-free zones, if their ancestors evolved where there were snakes (Coss & Goldthwaite, 1995; Towers & Coss, 1990). And the presence of snake and spider phobias suggests that humans could have similar templates for fear of snakes and spiders (Öhman & Mineka, 2001).

It's fairly easy to imagine the evolution of something like a template for snakes and spiders for several reasons: They are often dangerous to humans, tended to recur through wide swaths of space and time in human ancestral environments, and—importantly—spiders and snakes respectively share prototypical shapes and features that make them relatively easily discriminable. But this is not true of many other dangerous animals that inhabited human ancestral environments. They were quite variable in space and time and in properties like shape, size, and behavior, ranging from crocodiles to hippos to polar bears.

Crucially, while dangerousness might be somewhat predictable through simple cues like size, extremely fitness-useful knowledge about local danger is stored in the minds of other humans. This can greatly improve on the accuracy you'd get from cues like size alone or any combination of prepared cues (try guessing how dangerous a platypus is just from looking at it). One might expect, then, a prepared learning system to exploit culturally stored knowledge about which animals are dangerous and which are not.

Anthropologist James Broesch and I tested this idea in a study with young children in Los Angeles and among the Shuar of the Ecuadorian Amazon, where I conduct field work (Barrett & Broesch, 2012). Children saw flashcards of strange animals that they had never seen before (confirmed by a control group tested on the cards). They were told each animal's name, what it eats (plants or animals), and whether or not it is dangerous. Surprisingly, when tested five minutes later, children remembered which animals were dangerous and which were safe at about a 70% accuracy level, while the other information was at chance. Their performance was the same a week later, suggesting that the information had passed into long-term memory, and the same basic prepared learning effect was present in children from both cultures. This is especially interesting because children learned which animals were dangerous from a single, brief presentation. In a follow-up study with older children and adults in Fiji, we found evidence that danger learning biases may continue into adulthood, though the strength and nature of the effects change with age (Broesch et al., 2014).

These results are, I suspect, just the tip of the iceberg in terms of the landscape of what's easily remembered and what's not. Many naïve treatments of culture and socialization assume that humans are like sponges, absorbing whatever they are exposed to. This is a kind of cultural equipotentiality assumption: Like the assumption of behaviorist psychologists that any two stimuli could be associated equally easily, there is a widespread intuition that any cultural content is equally easy to acquire (Sperber & Hirschfeld, 2004). We don't really know if the landscape of cultural learning is flat or equipotential, and there are many tantalizing bits of evidence, some from research and some anecdotal, that it's not (*you* try telling your kids what are the right clothes to wear). There could be many hidden domains of prepared learning we don't yet know about, with their own input conditions and processing logics. For example, there could be prepared domains of learning that make bets about particular kinds of objects, like predators (Barrett, 2005a) or foods (Cashdan, 1994; Shutts et al., 2009), and that use particular kinds of cues to parse the world, such as using accents to parse people into social groups (Kinzler et al., 2007, 2009). There could also be even more complex designs, such as learning systems that decide, for particular kinds of information, whether it's better to acquire information from parents and other adults (e.g., generationally stable social or ecological knowledge) or one's similarly aged peers (e.g., rapidly changing trends such as personal style or age-group politics) (Harris, 1995).

Culture researchers have drawn an analogy between culture and the epidemiology of disease, where some diseases are more successful or catchy than others (Cavalli-Sforza & Feldman, 1981; Sperber, 1996). Biologist Richard Dawkins has introduced a similar (and itself very catchy) concept of memes on analogy to genes (Dawkins, 1976; Aunger, 2002; Blackmore, 1999). On the epidemiology view, some ideas might be "stickier" than others. They might be informational attractors, like magnets that draw attention, or viruses that insert themselves into memory (Claidière & Sperber, 2007; Heath et al., 2001; Lynch, 1996; Sperber, 1996; see also Henrich & Boyd, 2002). Information about danger seems to have this quality, but there are likely many dimensions of attraction (and perhaps repulsion). Advertisers, one suspects, already know quite a bit about this (Miller, 2009; Saad, 2007). It is likely to be an important area for future scientific work as well if we are to properly understand our own psychology, the dynamics of our own cultures, and why certain ideas spread and others do not. What is crucial in the evolution of such mechanisms (and what is very hard to examine empirically) is the cost-benefit structures of learning environments. If we could understand the underlying cost-benefit structures and how cultural transmission mechanisms adapt to them, we might go a long way toward understanding why people believe things that other people tell them, sometimes with no other evidence required than that other people believe it.

These examples reinforce the notion that cultural learning, like all learning, poses substantial problems of opacity. Learning only evolves because it increases fitness, but how do you know in advance that a particular piece of information will be useful? The opacity problem in cultural transmission is huge, and to the degree we can't solve it, we are prisoners of it, acquiring the bad along with the good because we are unable to distinguish which of the many things we could copy are the things that are actually fitness-good (Richerson & Boyd, 2005). There is no question that much of modern culture is filled with this useless bric-a-brac, the baggage of opacity. Our cultural learning mechanisms appear to have been designed to overshoot, learning arbitrary, useless things on the inductive bet that they might be useful; this is sometimes called overimitation (Lyons et al., 2007; Whiten et al., 2009).[3] But despite this, it is also true that any design features that *could* help solve the opacity problem—to narrow down the mappings between what we might acquire and what would actually be useful—would be heavily favored by selection.

As it turns out, there are many dimensions to opacity. In cultural learning, one thing that is opaque is the statistical association between a particular variant and fitness; this must be inferred. Another thing that is opaque is the *causal* relationship between the cultural variant and fitness: *how* it increases fitness. But there are other kinds of opacities as well. In language, we saw that the cultural variant to be acquired is something like "knowledge of language," including its rules, the meanings of its words, and so on. These must be inferred from uttered speech tokens, and so they are opaque. As it turns out, the language case is an instance of a kind of opacity that is common in cultural phenomena, where there is some overt production, or behavior, that is caused by an underlying knowledge structure. Any given instantiation of the behavior gives cues, or evidence, of the underlying knowledge, but does not reveal that knowledge itself, which must be inferred (I am not claiming that *all* culturally transmitted information is like this; some is, indeed, mere copying of surface cues). Is there anything that can help reduce the frame problem here?

Yes. In fact, we've seen several examples already. One is action parsers, which I discussed in chapter 4. The function of such parsers is to smooth over the infinite number of actual variations one might see over repeated instances of an act, such as reaching, to deliver a category judgment: This is reaching; this is an instance of a reaching event. Action parsers allow us to infer that two individuals doing different things at different times are actually both doing the same *kind* of thing: reaching. And, as I pointed out, our action parsers don't just carve the action stream at any old

[3] There is an analogy to error management here, and presumably overimitation could also be modeled using Bayesian priors (McKay & Efferson, 2010).

point, but at points related to goals. Reaching, for example, implies a goal: to grasp a thing. The parsers, therefore, carve up the action stream in ways that are useful for both learning and generalization across events (reaching events) and that deliver outputs in a format that is useful for other inference mechanisms, such as action prediction mechanisms and theory of mind (X is reaching for Y, therefore, it is likely that X wants to grasp Y).

To see why this matters for cultural transmission, think about how you might learn to tie your shoe. It could either be through explicit instruction—pedagogy—or through passive observation of others tying their shoes.[4] Either way, in order to learn it, you have to infer things like *first you grasp the end of the shoelace.* This means looking at other peoples' behavior and parsing it appropriately as a reaching/grasping event. This is not to say that no learning at all could occur without such a parser. But it should be uncontroversial that such parsers would give learning an enormous boost, and that humans probably do have them in some form.

Did action parsers predate the evolution of cultural transmission, or were they selected for because they improved it? Ultimately, this is an empirical question, and the answer could go either way (or perhaps both ways). Based on the arguments I gave in chapter 4, I think that at least some kinds of action parsers are likely to be phylogenetically old, far predating the evolution of the kinds of cultural transmission that humans have. Dogs, for example, probably have a general category of tooth-baring events. In primates, event parsers for things like grasping are likely to have appeared around the time that grasping hands appeared, because they would be useful for strategically modifying one's own behavior in response to others in contexts such as food competition. Knowing what another monkey is trying to do before he achieves it can be very useful, strategically, compared to having to wait until the fruit is already taken. Evidence that macaques have neurons (sometimes called mirror neurons) that respond to the category of reaching/grasping events suggests that monkeys do generalize across such events done by the self or others (Gallese et al., 1996). From an additivist point of view, this is good news for the evolution of cultural transmission mechanisms, because when design features such as prestige bias appeared, they could leverage the outputs of preexisting parsing mechanisms. However, it doesn't mean that additional modifications to parsing systems, including the addition of abilities to acquire new action parsing types, couldn't have been selected for once cultural evolution started to take off.

[4] There is a debate going on in anthropology, psychology, and biology about just how common each of these forms of cultural transmission is (see, e.g., Csibra & Gergely, 2011; Kline, 2014; Lancy, 1996; Thornton & Raihani, 2008).

Another kind of opacity-reduction advantage would be the ability to link actions to goal states: desired outcomes. This too requires, or is greatly helped by, an ability to correctly parse what the desired end state is, and then to infer the connection between actions and that state. For example, there is a difference between realizing that the goal state of shoelace-tying is a properly tied shoe with a certain kind of knot, and not just any old jumble of shoelaces. Even worse would be to infer that the point of the ritual is just to touch the shoelaces, move them around, and then remove one's hands. In the case of watching another individual crack a nut, a special mechanism might be required to infer that the goal is to remove the nut from the shell *with* the rock; without that, you might just see grasping events involving nuts and rocks and a shell-less nut appearing at the end, which could lead you to try all kinds of behaviors that don't combine rock and nut in the right way.

As I mentioned in chapter 1, children and adults seem particularly good at parsing these desired end states properly, and a host of experimental demonstrations show that when kids imitate, they attempt to produce what they have inferred to be the goal (Carpenter et al., 2005; Gergely et al., 2002; Lyons et al., 2007; Meltzoff, 1995; Whiten et al., 2009).[5] This can be either the desired end state or just producing an action in a certain way, like a correct ballet move; from a goal perspective, these are not distinct. Either way, if a learner is attempting to reproduce the underlying goal—if they are aware that there is an underlying goal that the person is trying to achieve and that not every single idiosyncratic surface feature of a particular token of an act might be relevant to it—this means that there will be many ways that they are not exactly copying what the other person does. Because of opacity, of course, there can and will be many mistakes even if one is *trying* to reproduce a goal, and the literature on overimitation shows that, often, kids will adopt seemingly arbitrary steps they have observed because they haven't been able to infer that they are unnecessary (Lyons et al., 2007). This reveals an additional design feature of these systems, namely, that they err on the side of caution, throwing in unnecessary steps because they *might* be necessary—possibly an adaptation to opacity (Gergely & Csibra, 2006; Whiten et al., 2009). Overimitation doesn't always occur, of course,

[5] Note, of course, that goals themselves are opaque. Many studies of this kind are designed to make the goal easily inferable by children. However, while behaviors may exist because they lead to certain outcomes, this may often be difficult or impossible to discern for the actor herself, let alone an observer. Additionally, of course, behaviors have multiple "reasons." The reasons why people pray, for example, include their conscious goals and intentions with respect to prayer, but there are presumably reasons why this behavior exists that those engaging in it don't understand—a distinction anthropologists sometimes make using the terms "emic" and "etic." Goal-based imitation depends on a match between the inductive bets of the imitation system and what the behavior is designed to achieve, usually in terms of the behaver's own representations (e.g., someone praying is trying to communicate with God).

and experimental demonstrations show that children can ignore unnecessary steps when they know what the right thing to do is (DiYanni & Kelemen, 2008). The point is that mechanisms that solve opacity problems can have large selective advantages in social learning.

Again, we can ask: Did the mechanisms that allow observers to associate parsed action types (e.g., reaching, tying) with goal states (e.g., properly tied shoes) evolve *because* of cultural transmission, or did cultural transmission evolve to exploit them, getting a boost because those mechanisms already existed to solve frame problems? In general, I suspect the latter, because these mechanisms would be useful for action prediction in species that didn't do any cultural imitation at all. However, there are likely to be at least some mechanisms that reduce frame problems specifically for the purpose of cultural transmission and cultural learning.

---

In Quine's "gavagai" scenario that we discussed in chapter 1, the adaptive problem facing the learner is to figure out the meaning of a word—a classic example of an opacity problem in cultural transmission. A variety of proposals and models exist for how children overcome the infinity of possible meanings to reach an inductive inference, including theory of mind (Bloom, 2000), heuristic rules such as the "whole object" assumption (Markman, 1989, 1990), and Bayesian models in which the learner simultaneously evaluates multiple hypotheses about the scope of possible meanings (Xu & Tenenbaum, 2007). Some or all of these mechanisms could well be operating, and they all have something in common: the inductive bet that when the person points to the rabbit and says "gavagai!" they are trying to communicate something. Because we're humans and equipped to do so, this seems like an obvious assumption to make. But it turns out to be immensely important and probably not obvious at all in the absence of specialized machinery.

Consider the following case of cultural learning: I catch a child's attention, say "Look! Can you do this?" and I tie my shoe. Or I start to play a videogame, or hit a golf ball, or cook dinner, or build a house. What is the child to make of the situation?

All of these are cases of what is known as pedagogy—the purposeful transmission of information to improve someone else's skills.[6] Developmental psychologists György Gergely and Gergely Csibra have suggested that humans possess

[6] Pedagogy, of course, is a synonym for teaching. Teaching is a special case of communication, which can be glossed as the purposeful transmission of information. Sometimes when I say "gavagai" I'm just trying to *tell* you something, and other times I might be trying to *teach* you something. What the difference is between these situations is not entirely clear, and a matter of debate. Gergely and Csibra have a particular hypothesis about pedagogy, which stipulates that it is specifically for transmission of information about *kinds*, rather than, for example, individuals or specific events (Csibra & Gergely, 2011). One could, of course, consider other

adaptations for pedagogy that help to narrow the frame problem in information transmission by cueing both parties to the fact that information is being intentionally or purposefully transmitted—in essence, that a demonstration is in progress. These mechanisms make inductive bets about how people behave when their behavior is pedagogical, as opposed to when they are doing something by accident or for non-communicative reasons (Gergely & Csibra, 2006, 2011; Gergely et al., 2007).

Gergely and Csibra propose that an important part of the pedagogy adaptation is a set of activation conditions or cues that tell the learner (usually but not always a child) that a teaching event is taking place. One source of such cues is what developmental psychologists sometimes call "motherese," infant-directed speech, or baby talk—a specific tone of voice, directed toward children, that can convey basic intentions on the part of the teacher, such as "pay attention to what I'm doing" (Bryant & Barrett, 2007; Fernald & Kuhl, 1987; Fernald et al., 1989; Papoušek et al., 1990). Others include cues like engaging the other person's gaze and then moving toward a third object—sometimes called shared attention or triadic awareness—and even so-called ostensive cues such as pointing (Liszkowski & Tomasello, 2011; Tomasello et al., 2005).

According to Gergely and Csibra's proposal, even young babies are equipped with detectors for such social cues and a motivational system that causes them to attend to the ensuing pedagogical demonstration and to attempt to learn from it, including, in some cases, imitating what they infer is the intended behavior. Moreover, they are motivated to make certain inferences based on the demonstration, such as that the knowledge being conveyed is generalizable and not specific to that particular event—a generalizability or genericity assumption (Csibra & Gergely, 2009). For example, children should assume that "gavagai" will mean the same thing to people other than the person who taught it to them.

Gergely, Csibra, and colleagues have conducted a variety of experiments that lend support to their proposal, demonstrating, for example, that giving or not giving pedagogical cues can have large effects on what children take away from the learning experience. We saw in the light box study in chapter 4 that children tend to imitate what they infer to be the experimenter's goal (e.g., turning on the light) without (necessarily) copying their exact behavior to reach that goal. A follow-up study provides evidence that this effect is much stronger when children believe the behavior is a demonstration for them (Király et al., 2013). When children received cues that pedagogy was occurring ("look!"), they attended to the availability of the actor's hands

cases of purposeful information transmission teaching, as in the case of ants that shape each others' paths toward a food source (Franks & Richardson, 2006). For the present discussion, I'll leave open the question of how narrowly or broadly to define teaching.

for turning on the light, imitating with their foreheads only in cases when it looked like the actor was demonstrating with his forehead on purpose (i.e., when his hands were free). When there were no pedagogical cues (the actor was not looking at the baby and didn't say "look!"), not only did babies imitate less, but they did not show as strong a difference between the hands-free and hands-occupied condition, suggesting that they were not being prompted to make inferences about the behavior *as* a teaching event.

Inferences about what the other person is doing and why can give a big boost toward solving frame problems, but they do not completely solve them, and there are many mistakes. Even in Gergely and Csibra's experiments, for example, some children didn't seem to get that the goal was to turn the light on, but instead inferred it was something like touching the light box with some part of one's head; some children caressed the box with their cheeks or kissed it. We can all think of many mistakes that might ensue from not making the correct inferences about what the right thing to do is in a cultural setting. What is remarkable, given the number of free parameters, is that we ever converge on any shared behaviors or knowledge at all.

Of course, if the additivist account I'm advocating is right, this *has* to be the case. If they are to evolve, mechanisms can't be paralyzed by frame problems; they *must* deliver an answer even if it is sometimes wrong. A mechanism can only be selected for, right from the beginning, if it delivers answers that are right more often than wrong, weighted by the fitness consequences. Selection can then push the mechanisms here and there to try to increase the frequency of correct conclusions and decrease the frequency of incorrect ones. Although we are not even close to knowing all the details, it seems like there must be a large number of enabling adaptations underlying our impressive abilities of cultural transmission, and it is an open and important question what their enabling design features are.

---

At the moment, culture and evolved mechanisms are treated by many as an either/or proposition: A given behavior might be due to one or the other, but not both. To some extent, this is just the newest incarnation of the ancient nature/nurture dichotomy. It also, of course, has a lot to do with the historical particulars of the current academic landscape. While many have called for "getting past" the nature/nurture debate—because genes and environments interact—I hope I've convinced you that we're going to need to do more than just accept interactionism. We need to embrace it and pursue its implications. The reason, of course, is that the details matter: We want to know what interactions produce human psychology and behavior, and how.

It seems all too common that when there are two parties in a debate and one side is forced to concede—because either the force of data or the force of argument compels it—they attempt to make the necessary concessions while ceding as little ground as possible. When there are two parties and *both* must concede something, they each come a little bit toward the middle, but refuse to meet. In the case of culture versus evolved mechanisms, there is still such a gap. Those on the culture side often concede that "some" evolved architecture is necessary to account for human behavior—some learning mechanisms, some cultural transmission mechanisms, and "of course" basic mechanisms of perception and emotion—but they prefer to imagine as minimal a toolkit as they can. Those on the evolved mechanisms side often concede that "of course" culture exists, but they treat it as *just* a kind of mechanism for calibrating to local environments, without recognizing that once evolved mechanisms begin to enable culture, the evolutionary dynamics of the species are (or can be) altered in major ways.

What I've tried to convince you of, at least, is that the story is likely to be complicated on both sides. More ambitiously, I hope I've convinced you that thinking about culture and evolved mechanisms as distinct ingredients that can be combined, as in a recipe, is probably too simple as well. Evolved mechanisms are necessary to enable culture and make it adaptive, but then, the culture actually enabled by those mechanisms is what creates the fitness benefits that drive them forward; neither ingredient exists or can be understood on its own. Dynamical systems can have interacting components, but those components only exhibit their dynamical behavior in the presence of the other.

In the case of human culture, once it takes on the properties of a Darwinian system and cultural elements become "replicators" (to use Dawkins's somewhat contentious term), then it becomes more than just an environmental calibration system (Godfrey-Smith, 2009; see Dickins & Rahman, 2012, for a contrary view). The dynamics can no longer be understood just in terms of games against nature. Among other things, as in genetic evolution, cultural evolution can now create designs. These, in turn, feed back on evolved mechanisms, shaping their inductive bets. We saw one case study where we are only just beginning to understand these dynamics: language. Let's turn to a few more cases in which our minds allow us to make things that, in turn, shape us.

# 10

## ACCUMULATION

The products of culture that surround us and inhabit us are so diverse and so seemingly limitless that it seems hard to imagine there is any logic to them, any predictability. Look around you: Would the details of what you see be guaranteed to occur in any possible universe? Would it contain Nintendo, breakdancing, neckties, *The Real Housewives of Orange County*, jeans shorts, handshakes, clowns, and moustache wax?[1] If you rewound the tape of history and let it play again, is this exactly what you'd see every time? And even with the benefit of hindsight, would any theory you can imagine actually be able to *predict* these things?

In evolution, as for any historical process, prediction is a tall order. You'd be as hard-pressed to predict the things above as you would be to predict the current position of the continents based on the starting state of the earth, or to sit in your armchair in 1,000 A.D. and predict World War I. But that isn't to say that these things aren't caused or that there aren't principles behind them. There are reasons why the events of World War I occurred, some of which we can understand. Similarly, there are reasons why humans transmit and learn information from each other and principles behind the evolution of cultural products. The products of culture include not just physical artifacts like tools and buildings and cornfields, but ideas, texts, customs, rules, performances, and techniques, from Beethoven's Ninth Symphony to matrix algebra to the speed limit on my street. The products of culture are in some ways part of our environments: I sit on chairs and drink coffee. And they are in some ways part of our phenotypes: I know the melody of Beethoven's Ninth and often obey the speed limit.

Culture exists because it structures the world and us. More importantly, it structures our interactions with the world and with each other. Cultural products evolve because of the roles they play in these interactions and, in particular, their fitness roles. Just like biological adaptations (though the distinction that this implies is, of course, ultimately false), the roles these products play are their functions. And the virtue of "ratcheting" cultural evolution is that it can climb complexity hills, creating

[1] Guess when this book was written.

cultural products that are intricately functional, like televisions, computer programs, and (sometimes) judicial systems.

Just as for the products of biological evolution, it would be a mistake to say that *everything* that is the product of cultural evolution can be explained in terms of its functionality; far from it. However, when thinking about the products of culture, it's important to adhere to a biological view of functionality, not an intuitive one. For example, we might balk at the idea that certain aesthetic features of an artifact, such as the three-inch high heels on a pair of shoes or the stripes on a tie—or even the tie itself—are functional. But that doesn't mean that these don't have a function in the biological sense. If those features play a role in perpetuating their own reproduction—as, arguably, the aesthetic properties of clothing do—then they *are* functional, even though they might not seem it. And just as every individual token of a biological adaptation need not increase its bearer's fitness, neither do, necessarily, individual cultural products. But the only explanation for culture writ large and the mechanisms that enable and sustain it is that they've taken us up fitness hills. The mountains of cultural detritus that surround us are, after all, extraordinarily costly in time, effort, energy, materials, and lives. In evolutionary terms, the benefits of all this stuff must have outweighed the costs.

In this chapter, we will consider the consequences of the fact that cultural evolution climbs hills, accumulating complexity. Some would disagree with this assessment, pointing to the arguably mistaken notion of cultural "progress." Indeed, some phenomena, like *The Real Housewives of Orange County*, would seem to argue against the idea that culture is going uphill.[2] But when taken seriously, the idea of cumulative cultural evolution seems hard to escape. What else, after all, explains the computer on which I'm writing this? What explains the fact that I'm using Windows XP, and that this fact will probably seem laughable by the time you're reading it? And if it's true that culture accumulates design *by* design, what are the consequences for the mental mechanisms that make it possible?

---

It is an interesting fact about adaptations that they solve problems even though the organisms possessing the adaptations don't need to know either *that* these problems

[2] Note that it's just as much a myth that cultural evolution inevitably leads to progress as it is that genetic evolution does. Biological evolution doesn't always lead to increasing complexity, nor does cultural evolution—and when increasing complexity does occur, it isn't always adaptive (McShea & Brandon, 2010). But when design does get more functionally complex, it is because of cumulative evolutionary change. There is no other way to explain the complex design of a mouse, either the kind that eats cheese or the kind that controls your computer.

are being solved or *how* they are being solved. For example, the immune systems of humans were busy solving problems of pathogen infection well before doctors even discovered they existed. And for other adaptations, like our eyes, we know that we have them and that they enable us to see, but we don't (necessarily) know *how* they do it. Moreover, we don't need to in order for them to work. As it turns out, the same is true of the cultural adaptations we acquire and possess (Richerson & Boyd, 2005). For example, the languages we use contain culturally evolved solutions to communication problems, including linguistic conventions such as grammatical rules and sound-meaning mappings. We can rely on these every day without having any idea how they work, or even that they exist.

This leveraging of culturally evolved design, without knowing how or why the design works or even that we're leveraging it, runs through our interactions with the stuff of culture, from language, to social norms, to human-made artifacts. My favorite example of this principle comes from a hallucinogenic drink prepared and used widely in the Amazon. This drink, called *ayahuasca* in Quechua and *natém* in Shuar, is made by boiling together two plants, *Banisteriopsis caapi* and *Psychotria viridis*. One plant (*P. viridis*) contains chemicals called monoamines, including dimethyltryptamine, or DMT, that produce hallucinations. The other plant (*B. caapi*) contains monoamine oxidase inhibitors, or MAOIs, such as harmine and harmaline, which prevent the monoamines from being rapidly broken down. By itself, DMT is orally inactive, probably because of rapid breakdown by monoamine oxidase, so without both plants, you'd get little or no hallucination (Riba et al., 2003). Most people who use this concoction know only *that* you have to put the two plants together, but they don't know *why*, either at the chemical level or even in terms of what would happen if you left one out.

This is a different kind of opacity: the opacity of the causal mechanisms whereby cultural adaptations provide solutions to problems. When you learn to make *ayahuasca*, you learn which plants you need to combine and how to combine them; but you don't learn *why.* What does this mean for the evolution of cultural learning and its design features?

You might think that such opacities would create a staggering learning problem, such that cultural phenomena like *ayahuasca* can't be learned or transmitted. But that is obviously not true, as the existence of *ayahuasca* demonstrates. Nor is *ayahuasca* particularly hard to learn to make; the recipe could be (and probably has been) written on a cocktail napkin. What's going on here, then? Didn't I claim that opacity problems are major obstacles to learning in species like capuchins that have trouble learning nut-cracking because they can't properly parse the behavior in terms of goals? If so, why isn't the *ayahuasca* opacity problem an insurmountable obstacle to humans?

The answer lies in thinking carefully about the underlying adaptive problems. If the hill-climbing model of culture is right, then *ayahuasca*-making evolved via trial and error over many generations of cultural time. It remains a mystery just how this happened, given the number of rainforest plants that could, in principle, have been tried before getting a combination that resulted in hallucinations. But what is certain, from ethnographic interviews of those who make it, is that once you've hit on the right combination, you don't need to know how it works in order to make it so that it works. All you need to do is effectively transmit the recipe. And the *ayahuasca* recipe was found via cultural evolution in a species that was already able to learn and transmit recipes: humans.

There are, in fact, two separate frame problems here, and they are solved differently. The frame problem of knowing which plants to combine was solved by cultural evolution via a walk through cultural possibility space. The frame problem of knowing what to learn when one is learning the recipe, on the other hand, is solved by evolved psychological mechanisms acting on cultural products that persisted because of their learnability. Solving the latter frame problem requires figuring out what the teacher is trying to transmit and which features the handed-down wisdom says are the crucial ones to keep in the recipe. No single individual knows why, but individuals do know *that* certain parts of the procedure are the "right" way, the way it was intended to be done. There can, of course, be many mistakes and the accumulation of useless details; but for functional artifacts like *ayahuasca*, the right steps must be preserved somewhere amid the unnecessary ones. One would expect, then, that an important problem that cultural transmission mechanisms need to solve is having some way of keeping the right aspects of the procedure from being corrupted, even when the transmitters don't know why they are right (for discussions of the problem of replication fidelity and the error threshold beyond which transmitted information begins to decay, see Claidière & Sperber, 2007, 2010; Dawkins, 1976; Eigen, 1971; Komarova et al., 2001; Sperber & Hirschfeld, 2004).

One way to do this is to have the "right" aspects coded as *goals* in the minds of successful learners and transmitters. For *ayahuasca*, the goal is to combine the two plants in a certain way, and successful learning means successfully inferring and acquiring that goal. The mechanisms proposed by Gergely and Csibra, as well as a host of others—theory of mind, action parsers, cause-effect inference systems—seem well suited to this. At least some of them might be required to be in place before cultural evolution of the ratchet-like kind can take off.

If cumulative cultural evolution provides us with a whole new category of solutions to problems—culturally evolved solutions with design features that are neither built into the minds of the people that transmit them nor would ever be hit upon by

single individuals in a lifetime of individual learning—then it's quite plausible that we have adaptations to exploit that fact. These would be adaptations that expect the world to contain culturally evolved solutions to problems, and instantiate inductive bets to exploit them. Language acquisition mechanisms, if they instantiate inductive bets such that languages will contain solutions to coordination problems like word order, would be an example of this. The case of *ayahuasca* suggests another possible example: mechanisms designed to enable us to learn about and use human-made artifacts.

---

Artifacts are a special kind of cultural object: They are physical things that we make, ranging from tools to houses, walls, roads, vehicles, and money (Margolis & Laurence, 2007; Searle, 1995). Some artifacts might not be physical things. Software programs and musical compositions, for example, have physical instantiations on computer hard drives and sheets of paper, but they can also be thought of in abstract, informational terms not reliant on any particular physical instantiation. A large and important category of artifacts, however, is physical objects: things that we make that *do* stuff, from keeping our drinks cold, to hammering in nails, to pleasing the eye. There are, of course, many accidental "artifacts," like smog and other forms of human-caused environmental change. But the category of things we typically think of as artifacts—things like tools that we make to interact with on purpose—generally have fitness benefits for us, and this is related to the fact that they usually have a use or function. In fact, they have culturally evolved to help us solve problems. This means that they have design features that make them fit for the functions they carry out. Of course, because artifacts are the products of cultural evolution, which has analogs of hill-climbing, drift, and the like, not *all* aspects of artifacts are functional, not by a long shot—but nearly all artifacts have design features that help them to carry out functions related to human goals (Petroski, 1992; Sperber, 2007).

On analogy to biological adaptations, this means that artifacts have proper domains and proper functions—the set of circumstances that shaped their design through a process of cultural evolution. Knives are for cutting, pencils are for writing, hammers are for hammering, and so on—not necessarily a single function per artifact, but functions are why artifacts have design features. This, in turn, means that the proper use of artifacts involves a special kind of convention: *This is how you use it*. While the functions and types of artifacts are diverse—paintings are to be looked at, roads are to be driven on, candy is to be eaten—perhaps the prototypical case of artifacts, and the earliest for which we have evidence in the archaeological

record, are tools.[3] And it's an interesting thing about tools that for each one, there tends to be a conventional or proper use, from cutting, to writing, to sitting, to drinking. In each case, it is this conventional function that has shaped the cultural evolution of the artifact.

Given this, you might expect that if there are specialized mechanisms for learning about artifacts, they will be designed to exploit the existence of such conventions. There will be "right" and "wrong" ways to use things, and the learning system will use cues to figure out what those are, thereby parsing the world into kinds of artifacts and building knowledge structures for using them.

As it turns out, there is evidence that humans are sensitive to the existence of such conventions in both how they learn about artifacts and how they represent them in memory. We saw some hint of this already in Gergely and Csibra's experiments with the light box: Most children inferred that the box was "for" generating light, and the large dome-like button on the front was "for" turning it on (Gergely et al., 2002). That the inferred function or purpose wasn't necessarily "something you touch with your forehead" was evidenced by the hands-occupied condition where the babies were happy to use their hands to turn it on, having inferred that the point was to turn it on by pressing the button.

There are many demonstrations that show that children infer the conventional use of an object based on very minimal information—sometimes only seeing the object used once. For example, psychologist Hannes Rakoczy and colleagues conducted studies where they showed young children a novel object, used it to push objects around a table in a certain way, and said, "This is daxing." Children strenuously objected when someone new came along (in the experiment, a puppet) and used the object in a "wrong" (i.e., unconventional) way. Some children tried to intervene, insisting to both the puppet and the experimenter: "That's not daxing!!! *This* is daxing!!!" Of course, you could argue that this doesn't prove that children associated the convention with the object itself, rather than the "game" of daxing, but the larger point stands: Children assume there are right ways to do things with stuff. These are conventionally (normatively) defined, and children are sensitive to this; they expect conventions about object use to exist in the world, and they have motivations with

[3] Interestingly, the properties of tools and their mental representation that I describe here—that there is a mapping between tools and conventional uses, with different tools having different functions—might not have been true of the first tools. Archaeologists debate their functions, and there was much less diversity in tool types than we see today. Indeed, on some accounts, there *were* no tool types in the very first stone tools. The explosion in tool types in complexity really didn't begin to take off until several hundred thousand years ago in Africa (McBrearty & Brooks, 2000). Adaptations for organizing tool representations according to conventional functions, then, probably have relatively recent evolutionary origins and have coevolved with the ratcheting complexity of tools in the world.

regard to those conventions, getting upset when they aren't followed (Rakoczy, 2008; Rakoczy et al., 2009).

What cues could children possibly use to infer the conventional use of an object? In worlds where artifact types evolve via cultural evolution to solve specific problems, a very useful generality holds: When people are using or interacting with these objects on purpose, they will tend to be using them in the conventional way (statistically at least). This means that children can use cues to intention and goal to disambiguate alternatives they might entertain about the conventional use. They can also use all of the other cognitive equipment at their disposal—action parsers, causal reasoning, inferences about others' beliefs, motivations, and expertise—to infer the conventional function of an object.

A host of studies has shown that the intentional/accidental distinction is important for inferring an actor's goal, which can in turn be used to infer what the actor is trying to do with the object, and therefore, in most cases, its conventional use (Behne et al., 2005; Carpenter et al., 1998; Casler & Kelemen, 2005; Kelemen, 1999; Kelemen & Carey, 2007). Imagine, for example, seeing someone use a hammer to pound nails: The person does it again and again, and if done properly, shows cues of satisfaction (or at least, an absence of cues of frustration). If the person hits his finger, he yells "ouch." If he knocks the hammer off the table by accident, he picks it back up. All of these things can be and are used to home in on the conventional use of the object. And in pedagogical situations, there are even more cues: When the person is trying to show you how the object should be used, you would expect them to use pedagogical cues to narrow down the possibilities even more ("no, not like that—like *this*!").

An alternative to making inferences based on how a person is interacting with an object is to make inferences from the properties of the object itself, based on what the psychologist James Gibson called affordances (Gibson, 1979): properties of the object that are well suited for some purposes but not necessarily for others and which, in the case of culturally evolved artifacts, will often (but not always) be culturally evolved design features.[4] Both human adults and children—as well as other animals—can and do make inferences based on affordances, such as shape, material composition, and other physical properties (DiYanni & Kelemen, 2008; Hauser, 1997; Hunt & Gray, 2004; Kelemen, 1999; Kemler Nelson, 1999; Kemler Nelson et al., 2000; Landau et al., 1998; Truxaw et al., 2006). But most cognitive psychologists think this is the hard way to do it when compared to the rich cues given by watching an expert, someone already familiar with the object, using it in its conventional way. Consider going into

[4] On analogy to proper and actual domains of adaptations, what an artifact was designed to do delineates its proper domain, whereas its actual domain—the set of things it can actually potentially do—is determined by its affordances.

a machine shop and trying to make inferences about the tools when nobody is there; it's possible, but it requires some fairly sophisticated knowledge of design principles, and moreover, you'll often be wrong. Now imagine the same thing, but you get to see the owner of the tools using them: much easier. And this is not necessarily because one frame problem (inferring affordances or inferring intent) is necessarily harder than the other or in the direction you might think. It's because the person using the tool is giving direct evidence of how it should be used.

As expected from these considerations, studies have shown that although children can make inferences from affordances, information gleaned from watching a person use an object tends to trump the affordance inferences the observer might have made, reflecting an inductive bet that information from the cultural source is much more reliable—though evidence that the person is using the tool incorrectly can moderate this (Casler & Kelemen, 2005, 2007; DiYanni & Kelemen, 2008; Want & Harris, 2001). Even more subtly, children take into account the relative degree of expertise of the person modeling the artifact use, imitating more precisely a person whose prior acts were observed to be efficacious and inferring the conventional use implied by the expert's behavior more than that of the novice (Birch et al., 2008; Rakoczy, 2008; Williamson et al., 2008).

There is a phenomenon called *functional fixedness* that appears to result from the mind's tendency to categorize artifacts based on their conventional uses, reflecting an ontological commitment that artifacts are "for" things (German & Defeyter, 2000; German & Johnson, 2002). Functional fixedness was first demonstrated by the psychologist Karl Duncker using the classic candle problem (Duncker, 1945). Functional fixedness studies generally involve two experimental conditions: one in which the conventional use is brought to mind or "primed"—not directly demonstrated, but implied—and another in which it isn't. In the candle problem, the conventional use of a box—to contain things—is primed by showing a box full of tacks. In the other condition, the tacks are shown next to the box, not in it. Subjects are asked to use the available materials to affix a candle to the wall. Duncker showed that subjects are slower to consider alternate uses of the box—tacking it to the wall to use as a platform to hold the candle—when its conventional function has been primed. Psychologist Tamsin German and colleagues have shown that functional fixedness appears early in childhood, but not immediately: Very young children don't yet seem captured by an object's conventional use (German & Defeyter, 2000). By about age six and continuing into adulthood, humans demonstrate functional fixedness in their problem-solving using objects with known functions (Adamson, 1952). German and I found that this occurs even in the Shuar, a society where people traditionally have used a relatively small number of tools to perform a variety of tasks (German & Barrett, 2005). If German is right, then functional fixedness may be the downside to

an adaptation that has an upside: Tagging objects with conventional functions means that as soon as we see an object and assign it to a functional category, we no longer have to use the more laborious process of making affordance inferences to know what to do with it.

Because of the process of cultural evolution, artifacts tend to come in *types*, of which any particular instance is but a token. This occurs precisely in order to reap the efficiency benefits of artifact conventions. If every token were different and had a slightly different function and different principles for using it, it would not be possible to generalize one's learning from one hammer to another, or from one instance of a word-processing program to the same program on another computer. Importantly, although there is variation in the shapes and sizes of hammers, they are all hammers, and they all share a common conventional function. Of course, artifact categories can be diverse and hodge-podge; there are roofing hammers, ball-peen hammers, etc. But if you buy a roofing hammer at the hardware store, it is a different token than *my* roofing hammer, and if you learn how to use a roofing hammer by using mine, you can inductively apply the knowledge to yours.

A feature of Csibra and Gergely's proposal about pedagogy is the prediction that children should possess a genericity assumption: They should assume that lessons derived from a specific instance should be generalizable to other instances (Csibra & Gergely, 2009). In the case of artifacts, it turns out that this is exactly what children and adults tend to do. Psychologists Adee Matan and Susan Carey showed that if children hear a story about an object, such as a hammer, and then about an individual who comes along and uses it in a nonconventional manner (i.e., as a hat), they still think it should be called a hammer, and that it still *is* a hammer (Matan & Carey, 2001). This is a strong intuition in adults as well, and it holds across cultures (Barrett et al., 2008). This is yet further evidence that people do not think of artifacts as "for" whatever they happen to be being used for in the moment, regardless of convention. Artifacts have an underlying purpose and a set of "right" ways to use them, and the learning task is not first and foremost to look at the object and infer from its affordance what you *could* do with it, but to try to figure out what you are *supposed* to do with it. This makes sense if the underlying learning mechanisms evolved in a cultural world.

Artifacts are a particularly interesting product of the human mind because they are likely to have a profound and ever-accelerating effect on human evolution. They are feeding back evolutionarily on the minds that created them, and the phenomenon of ratcheting culture, the ever-accelerating accumulation of complexity, can nowhere be better seen than in artifacts. Anthropologists generally agree that a major evolutionary milestone was passed when humans invented artifacts that could store information in ever-more complex form—the origin of written symbols, then language, then books, and now computers, television, and the like. But although these artifacts

are rapidly changing, each still obeys the basic rules of entities whose properties are transmitted through mechanisms of cultural transmission. In order to work properly, they must be used according to convention in the ways that they were designed. Even when learning to use something as complex as a word processor or fly an airplane, the learning task still comes down to learning how to use the thing in the way it was designed to be used—even though that design is the accumulated result of many, many years of cultural evolution spanning the minds of many, many individuals.

---

Even if you agree with me that our interactions with artifacts involve social conventions, you might wonder if this is particular to artifacts per se. Maybe conventions and norms are just a broader class of phenomena, and there are no mental adaptations specifically for artifacts. That's certainly possible, and it's an empirical question. There are some tantalizing hints of specializations for artifacts—for example, there are brain areas that appear to play a special role in our interactions with artifacts, activating stereotypical motor procedures when we see a familiar tool, for example (Culham & Valyear, 2006; Johnson-Frey, 2004). However, there is no question that conventions run throughout culture and human behavior. Another important and interesting class of conventions is what are sometimes called social norms; norms not for interacting with objects, but with each other, ranging from handshakes to marriage rules. Interestingly, many social norms have a feature that might reflect specialized design: They are moralized (Fessler, 2006; Kelly et al., 2007; Kohlberg, 1984; Krebs & Janicki, 2004; Sripada & Stich, 2006; Turiel, 1983). This implies a special kind of link with emotional and motivational systems: We want to do the "right" thing, and we get upset when others don't. If norms evolved to stabilize social interactions for fitness reasons, as evolutionary game theory suggests, then the fact that norms are moralized—even, perhaps, in cases where they seem rather arbitrary—this begins to make sense.

There is a debate about whether *all* social norms are moral. The psychologist Elliot Turiel and colleagues, for example, proposed that adults across cultures make a distinction between "moral" and "conventional" norms (Turiel, 1983). On this view, moral norms tend to involve harm and, Turiel proposed, they could be identified by the fact that people would not be culturally "relativist" about them, expecting moral norms to apply across cultures. "Conventional" norms, on the other hand, would not involve harm, but would rather be things like customs: the need to wear a suit and tie if you work at a law firm, for example. Turiel suggested that people would be relativistic about such conventions: They wouldn't expect, for example, that lawyers in Saudi Arabia would necessarily need to wear a suit in court. In support of this view,

a variety of studies have provided evidence that people do sometimes make this distinction, and that the distinction appears across cultures (Hollos et al., 1986; Nucci & Turiel, 1978; Smetana, 1981; Turiel, 1983).

However, philosopher Stephen Stich and his colleagues Dan Kelly and Dan Fessler have challenged the view that the moral/conventional distinction is quite so precise (Kelly et al., 2007). Indeed, on the view of conventions I'm advocating here, trying to divide social norms into "moral" versus "conventional" carries some of the same dangers as trying to divide linguistic conventions into the "functional" versus "arbitrary" ones. In a world where conventions regulate social behavior, violating them can indeed be fitness-bad for both the violator and sometimes others, raising questions about what one calls "harm." For this reason, Fessler has proposed that even seemingly arbitrary norms, like wearing a tie to work, might be moralized (Fessler, 2006). Hannes Rakoczy's studies of "daxing" support this idea, suggesting that even young children, exposed only briefly to a normative way to use a tool, get upset if someone uses it the wrong way (Rakoczy et al., 2009). And I'm sure we can all think of examples of seemingly arbitrary cultural procedures that get people all worked up if you violate them. Wearing a bathing suit to a solemn or formal occasion such as a funeral or a wedding could easily provoke moral outrage in some cultures, despite the fact that there are other cultures where people get married while being stark naked.

Note that Turiel's question concerns peoples' judgments about what is and isn't moral, which is not necessarily the same thing as what we, as scientists, should define as morality. If what we mean is something like cases where our behavior is shaped by what others are doing and there is a motivational/emotional system attached, then the scope of the moral could indeed be quite broad and perhaps structured into many sub-domains. Not surprisingly, then, there is currently something of a renaissance in the scientific study of human moral psychology, with a spectrum of views of what morality is, how we should understand it from an evolutionary point of view, and what underlying mechanisms might be involved. One currently popular question, as you might expect, is to what extent there are universals of morality, and how much of morality is variable across cultures (Haidt, 2007; Mikhail, 2011; Rai & Fiske, 2011). Another concerns the potential functions of moral judgments and intuitions, and the degree to which these are designed to help coordinate behavior, reduce harm, or create equity (Baumard et al., 2013; DeScioli & Kurzban, 2009; Gintis et al., 2008; Joyce, 2006; Nichols, 2004a). And there is the question of whether we should distinguish self-interest from morality, with many arguing that only cases of attitudes regarding third-party interactions in which you have no direct personal interest in the outcome should be regarded as moral, since otherwise, your attitudes just have to do with making others do what you want (DeScioli & Kurzban, 2009). On the third-party view,

morality might be specific to humans, because other animals, including chimps, get mad when people hurt them or take their stuff, but they don't generally stick their noses into others' business unless there is something to be gained or lost fitness-wise (e.g., intervening in a fight to help a genetic relative).

Here I won't weigh in on these issues except to note that I think the answer might include a little of all of the above. Personally, I don't think there are going to be easy natural lines to draw in most of these cases. As I'll argue in greater length later, I think the universals-versus-variation question is an important one, but it is often viewed simplistically; for example, if there are cases where intentional harm of another is allowed or even deemed morally good (e.g., corporal punishment of those judged to have done wrong), then some would say prohibitions against harm aren't a moral universal (Rai & Fiske, 2011). In one sense, this is true, just as it's true that not all languages have nouns and verbs, so these parts of speech are not linguistic universals (Evans & Levinson, 2009). However, all languages make the *distinction* between things and actions using various devices, and it's highly likely that all human cultures have strict rules regulating inter-individual harm, mostly (but not always) for the sake of reducing it (Mikhail, 2011). Similarly, just as it's dicey to try to tease out aspects of language that are "functional" versus "arbitrary"—because in a world of conventions, arbitrary things *are* or can be functional—it's dicey to try to judge how much of morality is simply a matter of conventions and how much involves a genuinely better or worse way to do something (for example, it's possible that driving on one side of the road is a little better than the other). And teasing out just those aspects of moral judgment that truly have no effects on your fitness (i.e., third-party interactions) might be hard too. Ultimately, moral judgment can evolve only if it *does* affect your fitness, so when peoples' moral buttons are pressed, there must in general be something "in it" for them; they are not truly disinterested parties.

Allowing, then, that there might be a spectrum of possibilities here, I want to briefly consider how the study of morality, norms, and social interaction more generally fits into the view of mental mechanisms I've been developing here. The view I've been advocating is that at heart, the function of all cognition is prediction in the service of action. As we've seen, psychological mechanisms can sometimes make explicit predictions—predictions that we're aware of such as "that dog is about to bite me"—but most of the time, they are making predictions in the form of inductive bets. This, I think, is how we should also view the mechanisms shaping our judgments and behavior in social interactions: They treat other people and social situations as things to be predicted, and make bets about the fitness-right thing to do. Sometimes that will be the moral thing, and if so, it's in our interest to do it (and sometimes perhaps—the dark side of moral judgment—it's in our interest not to, if we won't suffer consequences). What makes it fitness-right, of course, is whether the

felicity conditions of the inductive bet hold. In a game-theoretic world, these conditions could be something as simple as everyone here agrees that ties should be worn at weddings.

I've mentioned that in the domain of social interactions, so many mechanisms and processes are interacting that it's difficult to draw natural dividing lines. The same goes for the distinction between mechanisms involved in things like morality and cooperation, and mechanisms involved in social cognition more generally, including so-called theory of mind. Here I've advocated an expansive, functionalist view of theory of mind that is much broader than is typical in the literature. Where some restrict theory of mind to, in essence, the ability to reason about false beliefs, I've suggested that if it's about predicting others' behavior based on cues to their internal states, then it is really much broader, a spectrum of abilities that stretches across a wide range of species. In addition, on this view, theory of mind is likely to be deeply intertwined with morality, because the function of theory of mind is predicting the behavior of others, and so is—in the sense of prediction I described above—moral judgment. Cognitive scientist John Mikhail has suggested that we think about moral cognition as a form of causal cognition, and on this I agree (Mikhail, 2011; see also Cushman, 2008).

There are at least two ways in which thinking about others causally—as things whose behavior can be predicted—assists in interpersonal interaction. One is that we can adjust our behavior based on predicting what others will do, a form of Machiavellian intelligence. Another is that we can adjust our behavioral policies based on the effects those policies will have on others, including, shaping their future behavior. For example, one reason to get mad might be because it changes the probability that the person you're getting mad at will do whatever made you mad again in the future. If so, you'd predict our anger systems will take into account others as a causal system, getting mad as a function of the effects it's likely to cause in others (Sell et al., 2009). You'd also expect that our reactions to others' behavior take into account *why* that behavior occurred causally—which can entail taking into account both internal factors, such as others' intentions, and situational factors, such as the scope of actions that they could have done (Mikhail, 2011; Nichols, 2004b). Consistent with this, there is evidence for a variety of interesting interactions between theory of mind, causal cognition, and moral judgment: For example, doing something harmful on purpose is judged as worse than doing it by accident, and committing a bad act is, at least in some cases, worse than not doing something to prevent a bad outcome (Cushman, 2008; Young et al., 2007; Young & Saxe, 2009, 2011).

More broadly, given the importance of social interaction in human affairs, there is likely to be a complex dynamic of interaction between evolutionarily old and new mechanisms that structure our interactions with others, mediated by the products of

social evolution, such as norms and institutions: "things" like legal systems that alter the landscape of costs and benefits in social interaction, and that presumably install a variety of behavior-regulating representations in our minds. As we'll discuss in more detail in the next chapters, it can be risky here to try to tease apart what's evolutionarily new in human morality from what's evolutionarily old (i.e., what predates the human-chimpanzee split). The reason is that what's new, in evolution, is built on and from the old, and so new forms of social interactions, such as those mediated by culturally evolved norms, are likely to depend heavily on older mechanisms of social cognition like our abilities to make predictive guesses about others' behavior. Moreover, these older mechanisms shape the adaptive landscape in which newer mechanisms evolve, creating (potentially) interesting evolutionary dynamics. This will be important when we finally turn to the question of "human nature," and what such a term might mean.

---

In closing this discussion of culture and the ways in which what we shape also shapes us, there are several conclusions I'd like to emphasize.

One is that culture is not separate from the rest of our psychology or biology. I mean this in several ways. First and most obviously, humans are cultural beings and have been so for a long time. This means that we are likely adapted, in multiple ways, to being cultural. There are likely to be adaptations that enable us to produce culture, and these adaptations exist and are shaped because of the fitness effects of the culture they produce. Second, while it's currently popular to ask questions like "does this element X of our behavior result from culture or from our evolved psychology?" the nature of dynamical systems means, minimally, that it might not be an easy question to answer, and maximally, that it might not make any sense to ask it. To use a metaphor, suppose we release an indigo bunting and observe that it is able to use the stars to navigate home to its nest. Is the reason for this success to be found in the bunting's evolved psychology or in the stars? Clearly, it's both. However, the point I've been emphasizing is that it won't do to just say that: We must ask how and why the shape of the mind interacts with the shape of the world to produce this outcome. And the case of the bunting is likely to be much easier to answer than cases involving humans and culture because, after all, the bunting's navigational abilities have no effect on the stars: The arrow of causation points entirely in one direction. When it doesn't—as is the case with culture—it becomes much more difficult to place causation on one side of the arrow or the other. Critics of evolutionary psychology, of course, delight in pointing this out, but from it, they draw the erroneous conclusion that we should stop thinking about adaptations. Not only shouldn't we; if we want to

reach genuine understanding of culture as a biological phenomenon, as a product of the evolutionary process, we can't.

A final point concerns the importance of interaction. The case of culture and its products shows that interactions can't be ignored in understanding how things evolve: They evolve *because* of their interactions. If adaptations for language exist, for example, then they can't be understood except in terms of the languages they produce, nor vice-versa. More strongly, in an evolving dynamical system, we can't (usually) understand what's happening by imagining one component of the system as locked in place and the other adapting to it. To be sure, the details of the underlying dynamical processes, their timescales, and their causal properties all matter, but that's the point. As I described in the introduction, a kind of dualism currently holds sway in psychology in which the mind is populated by two kinds of things: the rigid stuff, sometimes known as modules, and the fluid stuff, known as domain-general processes. On this view, the rigid stuff is by definition non-interactive. Modules are isolated little cognitive reflexes, each doing their own thing automatically and without regard to what's going on elsewhere in the mind, and not adjusting themselves in the light of what's going on in the world. Instead, domain-general processes do that. The problem with this idea is that whatever adaptations in the mind might be, it is impossible to understand them except in terms of how they interact at multiple timescales from the momentary to the evolutionary within the larger system of parts of which the mind is composed.

PART V

# Architecture

# 11

## PARTS

Evolutionary psychology—not just what people think of it, but its true implications—hasn't yet fully percolated into the sciences of the mind. You can see this because people continue to insist that there are large chunks of the human mind that evolution doesn't or can't explain. And it is not a coincidence, I think, that these tend to be the parts of us that we hold most dear: things like our ability to reason, our creativity, our ability to respond to the here and now, our sense of self, our choosing of what we do and who we are. Many people who accept that natural selection and other processes acting over evolutionary time can explain *some* aspects of us, like what colors we can see or why ice cream tastes good, still insist that there is some category or class of mental phenomena that evolution can't explain. Sometimes, seemingly alternative explanations are favored: culture, socialization, individual experience. In some cases—this is particularly true when it comes to phenomena like consciousness—people are willing to concede that they don't know the explanation, but it's certainly *not* evolution by natural selection. What these cases have in common is a kind of explanatory walling-off: a willingness to concede that *this* category of mental phenomena (perception, basic emotions) is the result of psychological mechanisms shaped by evolutionary processes—but not *that* category (reasoning, flexibility, creativity, consciousness).

I'd like to return to the claim that I made at the beginning of the book: that this desire to wall off certain parts of our psyche from evolutionary explanation, and therefore banish them from study using the tools of evolutionary psychology, is wrong. There is not the evolved part of the mind and then the cultural part added on top like icing on a cake; culture and experience percolate throughout. And conversely, evolved mechanisms have to be part of the explanation for all mental functioning all the way up. In other words, I'd like to convince you that things like creativity, free will, and the ability of our minds to be shaped by experience at all levels are not only amenable to the same kinds of evolutionary functionalist reasoning I've been pursuing throughout the book, but are going to require it if we are going to fully understand those phenomena and why they exist in us.

In order to do that, I'm going to have to make a positive argument rather than just a negative one. I'm going to have to show you just how such an evolutionary account of what some psychologists might call the higher-order processes of the mind, such as those giving rise to reasoning, creativity, and consciousness, are possible. And to do this, I'm going to have to turn to the question of mental organization, or what is sometimes called mental architecture: how all the bits and pieces are assembled to create the whole of cognition (Carruthers, 2006).

Much of the groundwork for this, I hope, has already been established. Hopefully, we now agree that the set of possible evolutionary explanations of the individual bits and pieces and aspects of our psychology isn't necessarily as limited or narrow as it is sometimes depicted: For example, it's not just about what's present at birth or what is universal in the sense of being blueprint-identical in every person on earth. Hopefully, you concede that it's quite possible to have evolutionary explanations for mental processes that include both how they evolved and how they are constructed during development, via culture and individual experience and plasticity. And maybe we can agree that it might be useful to avoid, just for a while, the temptation to categorize every aspect of the phenotype into either innate or learned and talk instead in terms of evolved *design* and how it is built during development. Once we do this, we can have a much richer vocabulary for talking about adaptation that eliminates such paradoxes of novelty as why our visual system can see forks, why our mindreading system can parse television programs, and why we can learn to bake cakes, even though these are all evolutionarily novel. If you accept that evolution can and must play a role in explaining why we can do these things and can see at least in principle how a rigorous causal account could be made of the evolution of systems that do them—involving reaction norms, the statistical properties of environments, and the like—then much of the conceptual machinery is in place to build the kind of "all the way up" evolutionary psychology that I have in mind.

But not all of it. What remains to be confronted head-on is the issue of how all these elements, when put together, produce the whole of cognition. Given that we have a set of causal links that can get us from selection in the past to phenotypes in the present without any magical steps, can we be sure that when we connect the dots, there will not be anything magical left to explain? Or more minimally, can we see a pathway that can at least in principle squeeze out all the magical explanatory bits, thereby breaking down the wall between those parts of our cognition that can be explained by drawing a causal line back to evolutionary processes interacting with the world, and those parts that people insist must be explained in some other way?

I think we can. But in order to do so, we'll need to take very seriously the idea of interaction among parts and processes at many levels and scales. That is what the following three chapters of the book will be about. This chapter will argue that, contrary

to widespread claims in present-day psychology, human cognition is overwhelmingly likely to be achieved just as every other known functional process in organisms is achieved: via a functional division of labor among specialized parts that bite off manageable chunks of problems and interact by design to solve them. The next chapter will be about these interactions themselves, how they evolve, and how the parts are assembled to produce the whole, the very flexible, dynamic, and ever-shifting whole of cognition. Chapter 13 will turn to the questions of human nature and uniqueness: What, if anything, explains the gap that we like to think exists between us and other species, which we've so modestly recognized by calling ourselves *Homo sapiens*? And whatever it is, is it beyond the reach of the kind of evolutionary functionalist analysis I've been endorsing here?

The answer, as you might have guessed, will be at least a tentative no; tentative because we don't yet have adequate scientific theories of many aspects of higher-level cognition, evolutionary or otherwise. But this is partly, I'm arguing, because we've been afraid to look. Part of our problem as psychologists is that we've hobbled ourselves. While it's long been recognized that evolution by natural selection can explain things like perception, motor control, and our basic emotions, thinking about higher-level processes in terms of specialized design—that is, moving beyond the obvious point that they are adaptive to thinking about what design features enable them and why they exist in fitness terms—has been more or less taboo. If we're ever going to truly understand ourselves in evolutionary terms, then this is a taboo we're going to have to break.

---

The history of the debate over the fundamental parts or processes that make up thought is, to a first approximation, the history of psychology. It stretches back at least as far as Plato and Aristotle, who argued over faculties like emotion versus reasoning, to enlightenment philosophers like Descartes and Locke, who attempted to decompose the mind via armchair analysis into its fundamental components, to early psychologists like William James and Jean Piaget who began to gather data to test theories of the simpler, underlying mechanisms that they believed gave rise to the more complex whole of human thinking. It continues in modern psychology, enshrined in its most recent incarnation in the science of brain mapping. As our theories and understanding of brain functioning have become more sophisticated, so have debates about mind parts, but many of the underlying issues—are there really parts, are there lots or just a few, are they pre-specified, do they account for all of cognition or just some of it?—remain as fiercely contested as ever.

There are many ways of thinking about parts, components, and processes in the psychological literature. Open up just about any psychology journal and you'll see "boxologies," attempts to reduce larger phenomena to interactions of smaller parts. Perhaps the most explicit, notorious, and, some would say, pernicious concept of a mind part in psychology is the concept of a module, first formalized by philosopher Jerry Fodor in his 1983 book *The Modularity of Mind*. In this book, Fodor suggested that at least some mental processes might be modular, and he defined module in a very particular way as a kind of psychological Lego block: a tiny, innate, cookie-cutter-like psychological device, a reflex-like input-output machine that operates automatically and independently from other such machines, taking some narrow bit of informational input that it was designed to process, crunching the numbers, and spitting output for other systems to use. As part of this Lego-like view of modules, Fodor posited a variety of properties that he thought modules are likely to have. One was "encapsulation": They operate independently from other systems, which cannot access or influence their inner workings. Another was "automaticity": They operate like a reflex without any modulation from context or circumstance or conscious awareness. "Domain specificity" referred to the use of particular classes of information (and usually, in the way people now interpret domains, delineated narrowly and/or in terms of content—category specificity—as opposed to other informational properties). Yet another was "innateness": They develop the same way in everyone regardless of circumstance and do not depend on the features of inputs to shape their computational properties. At first glance, Fodor's idea seemed to map nicely onto the idea of an evolved psychological mechanism, and for this reason, evolutionary psychologists adopted the term "modularity" (Barrett & Kurzban, 2006; Sperber, 1994; Tooby & Cosmides, 1992). It soon became associated with evolutionary psychology to the later regret of Fodor and (at least some) evolutionary psychologists alike.[1]

Why regret? And why on both sides? The reasons might seem puzzling at first. Fodor's reasoning in *The Modularity of Mind*, while not explicitly evolutionary, was certainly functionalist. It tied the design features of modules to the functions they carried out in areas of cognition like vision or speech parsing. Encapsulation, for example, was regarded as a design feature that allowed each module to operate rapidly and efficiently without requiring inputs or control from other mechanisms. And innateness is viewed by many as going with an evolutionary view like bread with butter. So where is the problem?

[1] Note that Fodor was careful to stress that his notion of modularity was not all-or-none, but "ought to admit of degrees" (Fodor, 1983, p. 37). This is a point that seems to have been mostly forgotten in the subsequent literature. My comments here will mostly address those who think that modularity is a yes/no property.

The problem is that Fodor's account, and the way that modularity is typically envisioned in present-day psychology, is essentially dualist. That is to say, this view posits that the mind is composed of two kinds of things: the modular parts and the non-modular ones. This dualism is made most explicit in what are called dual-systems models of cognition (Evans, 2003; Greene, 2007; Kahneman, 2011; Sloman, 1996; Stanovich, 2004; Valdesolo & DeSteno, 2008). The terms for the two kinds in dual-systems models vary, but they are sometimes divided into System 1 and System 2. System 1 processes are narrow, automatic, inflexible, encapsulated, innate—in other words, modular. System 2 processes are broad (domain-general), non-automatic (open to control via conscious mental effort), unencapsulated (open to interaction, modulation, or control by other systems), flexible, and changeable by experience—in other words, non-modular. Following Fodor, dual-systems theorists explicitly associate the first, modular type of system with lower-level cognition—vision, speech parsing, and motor control are examples that Fodor used—and the second, non-modular type of system with higher- level cognition, such as reasoning, decision-making, and the like.

We can think of this as the "igloo" model of the mind: a crunchy outside composed of rigid innate modules and a soft center composed of general-purpose cognition, reminiscent of the Gary Larson cartoon in which a polar bear bites into an igloo and says to another bear: "Oh hey! I just love these things! Crunchy on the outside and a chewy center!" (Larson, 1984). The joke, of course, is that the chewy center is (presumably) a person. And indeed, that's how many people think about central systems: In the middle is where the self resides, the self-aware, conscious Wizard of Oz who presides over the automatic, mechanical modules.[2]

Dual-systems theorist Keith Stanovich has been perhaps most explicit in associating System 1 with evolved, specialized modules (he called them The Automated Set of Systems, or TASS) and System 2 with something else. Indeed, he has suggested

[2] As you might have noticed, there are at least two spatial metaphors here. In the igloo or central/peripheral model, the mind is like a sphere where the modules are on the outside (peripheral systems or input systems in Fodor's terminology) and the central systems are, of course, in the middle. In the higher-level/lower-level terminology that is common in psychology, lower-level processes tend to be the modular ones, and they are on the bottom. Lower-level processes include perception and emotion, which are easily associated with the intuitive idea of an instinct. Higher-level processes, of course, are on the top. These include reasoning, decision-making, and other processes intuitively associated with the self. Both of these metaphors are versions of a dualistic model of mind. Note also that the igloo metaphor has more ancient precedents, like Plato's allegory of the cave, where the "self" is trapped inside a cave and perceives the external world only through shadows projected on the wall. These are akin to representations produced by peripheral/input systems, which are held to be automatic and inaccessible to the self.

that we, as humans, might use our non-modular System 2 to escape the tyranny of our genes (Stanovich, 2004). While Fodor is much more hostile to evolutionary views (Fodor, 2000), he adopts a similar view of general-purpose, conscious processes as distinct from—and possibly reigning over—the modules: "If, in short, there is a community of computers living in my head, there had also better be somebody who is in charge; and, by God, it had better be me" (Fodor, 1998). Crunchy on the outside and a chewy center.

This is the view that I will call psychological dualism. As dualisms are wont to do, psychological dualism traffics in dichotomies: modular vs. non-modular, lower-level vs. higher-level, unconscious vs. conscious, automatic vs. controlled, innate vs. learned, specialized vs. unspecialized, and evolved vs. something else. Some of these distinctions might be real distinctions or at least dimensions along which mental processes vary, while some might not. What is odd isn't necessarily the dichotomies themselves. It's that in psychology, it's popular to think of all these dichotomies as essentially the *same* dichotomy. In other words, while evolved/not evolved could in principle be an orthogonal dimension to automatic/controlled and specialized/unspecialized, many if not most psychologists think the poles of these dichotomies align such that specialized mental adaptations all have one set of properties (innate, encapsulated, automatic, domain-narrow) and non-specialized processes have the opposite. The former are, of course, the lower-level processes like perception, and the latter are higher-level processes like thinking and deciding.

While dual-systems theorists do not explicitly deny that these second systems evolved, they resist the idea that they are composed of specialized adaptations whose form-function relationships can by analyzed in the same way as any other adaptation. This is because (so the dualist logic goes) they are, unlike modules, general-purpose, flexible, non-automatic, and use a broad range of information. In other words, higher-level systems are a different kind of beast, not amenable to analysis in terms of design features evolved in ancestral environments to solve adaptive problems.

Why is this wrong? It's not, as I have suggested, because all of the dichotomies or distinctions involved are necessarily wrong. It's fairly well established, for example, that some brain processes are more subject to conscious control than others. What is wrong is to assume that the mind contains two fundamentally different kinds of things: the specialized parts, which carry out specific functions and are automatic, encapsulated, etc., and the unspecialized parts, which are not. Instead, while it might be the case that the lower-level/higher-level continuum is real in some ways—for example, it's true that information first must pass through perceptual mechanisms before being useable by later processes—it is highly unlikely that the process of adaptation via natural selection doesn't account for how the higher-level stuff gets done too. In fact, nobody really disagrees that higher-level cognition is functional

and indeed adaptive, and that our higher-level mental skills have probably played an important role in human evolution. The resistance seems to be in thinking about higher-level processes *as composed of adaptations*.

Now, first law logic implies that there is no reason at all to think that any two adaptations need to have the same properties or design features (Barrett & Kurzban, 2006). Indeed, to the extent that two different adaptations have evolved to do different things, we *expect* them to have different design features. And when you look at Fodor's list of the properties of modules, you see that it is really a list of design features: Properties like domain-narrowness, encapsulation, and automaticity are properties that some cognitive processes might have but not all cognitive processes must, as dual-systems theorists themselves acknowledge by positing System 2 processes. In other words, cognitive processes can and do exist that do useful, adaptive things with information without being domain-narrow, encapsulated, or automatic. If such processes exist in the mind and are adaptive, they must, barring other alternatives, be the result of evolution by natural selection. This means that we ought to start thinking about them as adaptations, with all that this entails. It entails, of course, that we think about adaptive diversity: Different mechanisms in the mind carry out different functions and therefore will have different designs. But it also entails thinking about where adaptive flexibility comes from. Does it come from one or a few general-purpose flexibility mechanisms interacting with an army of rigid, inflexible ones? Or does flexibility in the mind arise, as it does in many other biological systems, from the interaction of many specialized parts?

---

Before we start thinking about what the parts of higher-level cognition might be like and how they might differ from the prototypical modules of lower-level cognition (and more trivially, whether we want to use the same word to refer to them both), let's first visit those lower-level modules. As I mentioned, Fodor's reasoning about modules in *The Modularity of Mind* was quite functionalist, even if it was not explicitly adaptationist. Fodor chose features that he thought would be associated with modules for a reason: They were the features that he thought made most sense given the main type of function that he thought modules carry out in the information-processing architecture of the mind. That function, in essence, was to serve as in input pipeline to the rest of the mind.

Here is something that Fodor was right about: Information-processing systems exist through which *all* information from the external world must pass if it is to get into the mind at all. The first stage of such systems, as I mentioned in chapter 1, is sometimes called sensory transduction, in which information is translated from its

external form—light waves in the case of vision, air pressure waves in the case of sound, mechanical contact and force in the case of touch, chemical traces in the case of smell—into the informational currency of the brain, namely, patterns of neural firing. What happens next is the stage that Fodor was most interested in, and the stage that he used—mistakenly, I'm arguing—as his prototype for thinking about all modules. It is the set of processes via which the patterns of neural firing delivered by sensory transducers are interpreted and packaged for the rest of the brain to use. This is the stage we conventionally call perception.

Fodor's reasoning about the properties of form-function fit that one might expect in perceptual mechanisms, or input systems as he called them, is compelling, and much of it continues to fit what we know about perception several decades later. His argument was that the primary aim of perceptual mechanisms is "veridicality": to deliver a "true" representation of the outside world based on the information delivered to them by sensory transducers. On this view, the primary function of perceptual mechanisms is reconstruction of the external world from the somewhat impoverished information presented by the eyes, ears, nose, mouth, fingers, etc. This latter bit perceptual psychologists generally agree on: The primary function of perceptual mechanisms is to reconstruct the world around us for the rest of our brain to operate on, and the information presented by our sensory organs is a paltry slice of what's actually available in the world. Consistent with Fodor's view, nearly everything about perception is still understood in terms of the adaptive problems inherent in making inferences about the external or distal world from the sensory or proximal information available (Palmer, 1999).

What's not exactly right is the idea that perceptual systems are entirely veridical: They do not in fact always tell us the truth, the whole truth, and nothing but the truth. This is because truth is not the ultimate arbiter of which variants of perceptual systems make it into the next generation; fitness is. More specifically, which perceptual systems do or don't last over evolutionary time depends on fitness differences between the variants that actually exist at any given time. This means that many "truthy" variants never appear (e.g., the ability to hear radio waves), and many truthy ones that were present in our ancestors, such as the ability to detect certain smells, have probably been lost in us as the costs and benefits of retaining them changed. And the perceptual mechanisms we do have sometimes gloss the world in various ways rather than give it to us straight, either because it's fitness-better to do that or because tradeoffs favor a shortcut over complete accuracy. An example is the perception of color, where an object appears to be the same color under different lighting conditions even though the wavelengths of light coming off of it differ, because the fitness-relevant adaptive problem is to track the same object across varying lighting

conditions, not to give us a percept of the true distribution of wavelengths reflected from the object (Shepard, 1992).

With those caveats, certain aspects of Fodor's proposal—in particular, the package of features he believed would be associated with modules—make sense for perceptual input systems, given where they sit in the architecture and their role as the mind's gatekeepers. In particular, because all downstream decision-making systems are relying exclusively on the outputs of these mechanisms—such as the systems that tell our arms and legs how to move when we see a snake—it makes fitness sense that they would be fast. And because these systems are number-crunchers that do things like calculate distances, find the edges of things, add color and shading and depth, it makes sense that they would apply similar algorithms across a wide array of contexts and situations, algorithmically and reflexively, without need for modulation by other systems or worrying about the content of what they are processing (e.g., whether the yellow object whose shading they are computing is a lemon or a tennis ball). For these purposes, properties like automaticity and encapsulation might well be favored by natural selection. And they are consistent with much empirical data on the early stages of perception: It does seem to be carried out by massive banks of tiny number-crunchers to which information is fed as if via a pipeline, processed automatically and reflexively—each gizmo in the bank biting off its bit of information and chewing it without consulting its neighbors or higher-ups—and then spitting out the results to the next level up in the hierarchy. There are, of course, exceptions, including top-down effects even in early stages of perceptual processing, where later systems' inferences about what is being processed modulate lower-level systems. Everything, even here, is subject to the first law. But broadly speaking, Fodor's model of the prototypical module is not a bad one for early stages of cognition, because it makes adaptive sense given the form-function fit relationships we might expect such mechanisms to have.

---

But do these properties make sense for all systems? In other words, in order to be an evolved psychological adaptation, do you need to be encapsulated, automatic, inflexible, and all the rest?

Almost certainly not. Indeed, the adaptive logic that makes sense of this package of features begins to break down almost as soon as we leave the earliest stages of cognition. More broadly, it breaks down as soon as we leave processing contexts in which a strict assembly-line design is best for getting the job done—which means it's probably not the best design for much, and perhaps even most, of the mind.

When I say assembly line, what I mean is a design in which each component in the system simply takes information automatically from the previous component, processes it, and hands it on to the next component in the chain, without any modulation on the basis of what it is processing, or any capacity to interact with other elements in the system other than the components immediately before and after it.[3] As designers of factories have discovered, this kind of design, like any specialized design, has both benefits and costs. The benefits come from the specialized division of labor among the processing units in the chain. But this design is efficient only under certain circumstances. In particular, it works best when the operations to be performed on the "product"—in this case, information—are the same in every case. Under those circumstances, it may well be optimal to assign each unit or person in the assembly line a relatively narrow, repetitive, automatic, reflex-like task and to have that individual do only that task over and over without having to consult his neighbor or modify his procedure depending on what it is he is processing. The cost to such a system, on the other hand, is that it is inflexible. It only works if conditions are set up such that what each person receives is prepared for him in just the right way by the last person in the line, so that he doesn't have to consider possibilities different from what he is currently doing.

Imagine, for example, that I work in the assembly line of a sausage factory, and my role in the assembly line is to remove meat from a meat grinder and stuff it into sausage casings. This allows me to specialize and focus on a particular task, and I can become quite skilled and fast at it. Much of the efficiency comes from the fact that I don't have to worry about what's coming out of the grinder. Whatever comes out, I take it and stuff it into the sausage casing. Presumably much of cognition is like this: for example, taking information off the retina, slicing, dicing, packaging it, and sending it off.

The benefit of having me take whatever comes out of the grinder and stuff it into the casings is that I don't have to worry about what's coming out of the grinder. But this, as owners of meat-processing plants have discovered the hard way, can also be the system's Achilles heel. There is a tradeoff such that the less time and effort I spend evaluating what's coming out of the grinder, the faster I can stuff it into the sausage casings; but I am also less likely to detect any possible problems. That's why designers of such factories add elements beside than the narrowly focused, automatic, encapsulated sausage-stuffers: higher-level supervisors and inspectors who watch what's going on and regulate the activity of the system. They can execute top-down

[3] I've also called this a "pipeline" model of cognition (Barrett, 2005b). Here I'm using the assembly line metaphor in order to focus on the computational work done by each unit in the chain.

decisions, shutting down the line when something undesired goes into the grinder, redirecting the activities of certain people on the assembly line, and so on.

It is well known that in cognition, such higher-level systems exist: systems that scrutinize the outputs of lower-level mechanisms and use these to control other aspects of the system's operation. A good example is the phenomenon known as attention (Itti & Koch, 2001; Knudsen, 2007; Treisman & Gelade, 1980). Although it might be the case that your low-level perceptual mechanisms automatically give you a picture of whatever you happen to be looking at in assembly-line fashion, there are clearly other mechanisms that decide where to direct your eyes, and even what you attend to in your immediate visual field (Scholl, 2001). And those attention-guiding mechanisms must be doing this—in fact, they must have been selected to do this—at least in part based on the *content* of what's being seen, such as whether it's an oncoming car or a passing cloud (New et al., 2007a).

Does this mean that the higher-level mechanisms that make inferences and decisions based on the outputs of perception aren't specialized? There is no reason to think so. But it does mean that they are likely to have different properties and different design features than the lower-level systems delivering the percepts they operate on. For example, in analogy to the sausage factory, it's likely that there are properties of information that high-level systems in human cognition attend to, such as whether the representation of object X being emitted by the object-parsing system is a stuffed bear or a real bear, even though the perceptual assembly lines filling in the color and shading information on the bear don't. And the whole purpose of these higher-level systems is to coordinate and adjust the behavior of other systems flexibly, depending on the information they receive: to run away if it's a real bear, and to hug it if it's not. The properties of these systems, then, must differ in some ways from Fodor's set of features: a little less automatic, a little more flexible, a little more interactive, a little broader in the scope of information they examine. And, given that these systems must do things like decide whether to look at a car or a bear or a menu, they can't use only innate information. More likely, these mechanisms are adjusted and shaped by experience—and presumably, this is by design.

That said, however, control via higher-level mechanisms is not the only way for systems composed of modular parts to achieve flexibility—far from it. In addition, flexibility can be achieved by relaxing one of Fodor's assumptions (or at least, what has been widely taken to be a necessary feature of modules), namely, that modules are unable to interact with their peers. If there are any lessons to be drawn from the growing field of self-organizing systems, it's that adaptive behavior can emerge from the interaction of many parts each of which operates according to its own logic (Holland, 1992). In some cases, these emergent patterns can be highly adaptive—especially in cases where the properties of the parts have been shaped by a history

of selection specifically in order to generate those emergent patterns, a phenomenon I will call *selected emergence* or *designed emergence*.

There are many emergence-type models of cognition in which higher-level organization emerges from the interaction of parts. Neural networks are a familiar example (Rumelhart et al., 1986–87). An older conceptual model of the same idea is Oliver Selfridge's pandemonium model (Selfridge & Neisser, 1960). This was originally proposed as a model of how recognition of written letters like A and B might occur. In this model, the perceptual process is entirely composed of modules, or "demons," which are arranged in a hierarchy such that modules at a given level interact in a manner analogous to voting to influence what happens at the next level up. At the lowest level, each demon scrutinizes the perceptual array for a given feature. One demon might look only for edges with a 40 degree angle to vertical, another for edges with a 45 degree angle, another for 50 degree angles. Some might look for curves, some for gradients of shading, and so on. Each is specialized and relatively narrow. And each emits an output—a "shout" in the original, metaphorical model—if it finds its feature in the input. These shouts vary in volume and are proportional to the certainty with which the feature has been detected, because the demons have "fuzzy" input conditions: They look for a match to a prototype and shout corresponding to the degree of match. Such fuzziness is thought to be a common feature of neural pattern-recognition systems. For example, a slanted line demon might emit a shout for A but not B, because A, but not B, contains slanted lines. A curve-detecting demon might shout only for B, because B, but not A, contains curves. These shouts then serve as vote-like inputs to the demons at the next level of the hierarchy, which represent combinations of features, like the presence of both right-slanted and left-slanted edges in the same letter (e.g., in the letters A and V, but not B or Z). The middle-level demons listen to the shouts of their colleagues below and then shout upward with a volume corresponding to the overall (emergent) intensity of the shouts corresponding to their inputs, and so on all the way up until a unique letter is identified in the form of an individual letter demon hearing enough shouts to satisfy him. The answer, therefore, is both probabilistic and emergent in that there can be varying degrees of uncertainty about the stimulus instantiated in the varying volumes of the shouts, and in that the answer is determined by the interactivity of the collective.

There have been debates about whether a pandemonium model is likely to be the correct model of perception (Grainger et al., 2008; Massaro, 1989). In all likelihood, *something* pandemonium-like is occurring, in the sense that perceptual processes combine simpler elements hierarchically into an overall percept (Crick & Koch, 2003; Dehaene, 2009; Palmer, 1999; Ullman, 2007; Yuille & Kersten, 2006). Whether or not the pandemonium metaphor is precisely correct, the point is that it illustrates how a system that is composed entirely of specialized parts, each of which executes its

narrow, domain-specific task, can produce a remarkable degree of flexibility when those components interact, and without anybody in charge. Flexibility through interaction is the whole point of their design. And, crucially, they are composed of parts with functions—all the way up.

---

Let us now consider in more detail how adaptive problems might change as a function of where mechanisms sit in the architecture of the mind (i.e., high or low level) and what they do. Broadly speaking, information-processing mechanisms face two kinds of adaptive problems: what we might call interface problems, or problems of how to get information and where to send it, and processing problems, or problems of what to do with information once they get it. As we'll see, these problems are intimately related, because problems of what to do with information can depend on, first, what other mechanisms you're getting information from and why, and second, for what reasons you're sending the information you've processed out to other systems. Since we have to pick one, let's start with processing problems: in other words, the question of how the form of a mechanism's information-processing operations is fit to its function.

Problems of how to process information are problems of algorithmic design and are therefore the kinds of problems that cognitive psychology typically deals with. For example, do categorization mechanisms operate by tallying up the features of items, weighting each one, and then making a judgment? Or do they operate via a fuzzy, graded comparison to a prototype (Rosch, 1978, Smith & Sloman, 1994)? Do the mechanisms that construct and parse words and sentences use discrete rules, or are they more probabilistic (McClelland & Patterson, 2002; Pinker & Ullman, 2002)? Do the mechanisms that trade off future options against current ones use a hyperbolic time-discounting function or a function with some other shape (Ainslie & Haslam, 1992; Read, 2001)?

All of these are questions about the nature of the functions or algorithms that information-processing mechanisms instantiate. It is widely agreed that because neural systems process information, they can be modeled as mathematical algorithms that take inputs in the form of neural information arriving at the mechanism in question, perform operations or transformations on those inputs (including, sometimes, leaving parts of the information alone), and generate outputs in the form of the transformed information, which is then sent along to other mechanisms or stored in memory. When thinking about the design features of mechanisms, we can think about the design of all three of these aspects—inputs, operations, and outputs—in adaptationist terms (i.e., in terms of fit).

As we have seen, mechanisms can have different input designs: different ways of receiving information from other mechanisms. One distinction that can be made is between mechanisms that select their own inputs versus mechanisms that have their inputs selected for them (Barrett, 2005b). The latter is illustrated by what I've been calling pipeline or assembly-line systems, where the sausage-stuffer has no choice over what he puts into the sausage casings. His inputs are determined entirely by what comes out of the meat grinder. He does, therefore, have inputs, and they are domain-specific; but he has no design features for solving the problem of choosing *among* inputs. Certainly, some cognitive mechanisms will be like this: in essence, hard-wired to the outputs of other systems. An example might be mechanisms that get information immediately from the retina. There is nowhere else they are going to get their information from.

Many (and perhaps most) mechanisms, however, will be designed to operate contingently, depending on what kind of input they encounter. This can include either deciding whether or not to process input that is sent to them, or more actively, searching through a database of information and deciding what to process. In either case, mechanisms will have input conditions, some set of criteria for deciding what and what not to process.

When I say deciding, this doesn't imply conscious choice. It can be done algorithmically, just as a coin-sorting machine sorts coins of different sizes into separate slots via a physically instantiated algorithm, an array of slots arranged in a particular spatial pattern. And the decision about whether or not to process an input needn't be entirely binary (either/or) though it can be. For example, in the pandemonium model, we encountered the idea of fuzzy input criteria, where the strength of processing corresponded to the degree of match to a prototype. Neural networks are particularly good at this, because they scrutinize inputs along many dimensions rather than just one. Processing depends on strength of activation along these many dimensions, either via summing them or in some more complex form of combining inputs (Rumelhart et al., 1986–87).

What we might expect is that mechanisms designed to use the outputs of other mechanisms will be sensitive to certain systematic features of those outputs: They will make inductive bets based on them. Earlier, I mentioned the concept of an informational tag that can be added to information by one mechanism, and that other mechanisms can later use. Examples we saw included tagging various perceptual information as belonging to the same object, tagging an object as animate or as a member of a more specific category, such as a lion, and tagging the perceptual representation of a face with the identity of the person whose face it is. In each of these cases, a mechanism or set of mechanisms has done some processing and produced output of a form such that later mechanisms won't have to do that same

processing; a classic example of division of labor. Downstream mechanisms then can use these tags in a variety of ways. For example, they can use the tags to solve the kinds of information-routing problems mentioned above: Is this something I am designed to process or not? And even once information has been admitted for processing, the mechanism can use these tags to modulate processing in adaptive ways. For example, facial expressions are probably tagged in some way according to both their emotion category (fear, anger) and intensity. Decision-making processes can then use these tags in their own computational procedures; for example, deciding whether the level of anger seen in the other person merits apology, flight, or some other strategy.

What is the range of forms that higher-level, non-Fodorian mechanisms—ones that go beyond merely reflexive, low-level processing of tiny bits of perceptual information—might take? Just as for developmental reaction norms, we can think of information-processing mechanisms as instantiating functions that transform inputs to outputs, and these functions could be as diverse as anything that can be done with neurons—essentially, an infinite palette. Ultimately, of course, these will be shaped by their effects on fitness, no matter how proximal or distal they are to actual behavior. Edge-detectors compute object boundaries, ultimately, because of the interactions they enable us to have with those objects. Mechanisms that encode memories make inductive bets that those memories will one day be useful, in the sense that the net fitness benefits of memories stored will outweigh the net fitness costs. Motivational systems that cause us to get bored with what we're doing now and shift our activity to something else must, ultimately, have their tipping points calibrated by fitness. These are all very different systems with different informational shapes, but all are shaped in the service of fitness.[4]

---

[4] Some would argue that I have focused excessively here on computation and the idea that the function of mental mechanisms is exclusively to transform inputs to outputs. For example, Van Gelder (1995) has suggested that dynamical models of mind are inherently non-computational. While there is room to disagree about this, I'd like to stress that the view of evolved mechanisms I'm advocating here is broad enough to include phenomena and models that are sometimes considered outside the scope of standard computational models, including dynamical systems models, neural network models, and embodiment models. Like all models, these are just different formalisms for approximating the causal properties of the real underlying stuff, which is biochemical in nature. It seems to me there is no inherent conflict between thinking of the mind as a system of neural networks, as a dynamical system, as a probabilistic computational system, and as an embodied system. The key point is that all can be thought of in terms of properties that have been shaped by the evolutionary process.

What does all this mean for thinking about the design features of higher-level cognitive systems? It means that none of the stereotypes we typically associate with evolved mechanisms are likely to apply, because those stereotypes come from areas of cognition like perception and motor control. The design of higher-level cognition is unlikely to look like the wiring diagram of a transistor radio, with a few narrow, automatic, inflexible Lego-like parts hard-wired together. It's more likely to look like the worldwide web or a cloud computer, involving flexibility and context-sensitivity down to its very core (Pinker, 2005).

But this doesn't mean that specialization isn't involved. Indeed, it's quite likely that specialization is going to be *required* in order to achieve the remarkable degree of flexibility we see in higher-level cognition (Carruthers, 2006; but see Machery, 2008). I say this because while it's true that flexibility and specialization are widely thought to be opposites in some kind of zero-sum tradeoff, there's actually little evidence from real material systems, such as biological or human-engineered devices, that this is true. Instead, what the engineering world shows us is that there is a tradeoff between being specialized and being generalized—but being general and being flexible are not the same thing. Being general-purpose simply means you have fewer design features relevant to any particular context. It does not mean you are able to adjust or adapt yourself to diverse contexts, *unless*—and this is an important caveat—you have design features to do so. For example, a roll-up measuring tape might be said to be flexible because it can adjust itself to the length required, but that's because it's designed to do just that. More broadly, as I've mentioned, the history of design of devices like computers and telephones has shown that increasing the flexibility of these devices has generally been a matter of adding more features, not taking them away. So-called smart phones are more flexible than the earlier generation of cell phones—they can do more things—because they have more software, not less. And these new features are designed to increase the flexibility of the system in just the ways desired, like software innovations that allow more passing of information between specialized subsystems (e.g., your browser sharing with your photo or text editor and then allowing the results to be reposted online; Baldwin & Clark, 2000; Kurzban, 2010).

If the arguments I've been making in this book are right, then the mind too is a system whose entire logic is the logic of passing information between systems in adaptive ways. It traffics in representations that flow through the mind, being modified and passed along by the mind's mechanisms. This means that those mechanisms *collaborate* to modify information, each modifying the information passing through it in its own function-specific ways, but leaving other modifications to other mechanisms. It also means that most mechanisms must be context-specific in their operations, modifying information when it is part of their job description but leaving it to

others when it is not. And if this is to produce selected emergence—computational outcomes that emerge from the interaction of parts by design—then different parts in the system must be designed to rely on work that other mechanisms in the system are doing or have done.

Let's consider an example that makes use of some of the abilities I have described in the book, and consider how such abilities might be instantiated in a system composed of collaborative but specialized parts. Imagine a situation that happens to all of us many times a day: you encounter someone who is making a facial expression, you interpret the expression, and then you decide how to react. Now, there is no question that humans exhibit a massive amount of flexibility in this regard, both within and across individuals. How you react depends on an enormous number of things. Obviously, it depends in part on what psychologists would call the basic category of the facial expression (e.g., angry, sad, happy, fearful). But it also depends on more subtle details of the expression: Is the person hysterically sobbing or just a little sad? Are they red-faced and screaming or just irked? Are they ecstatic like they've just won the lottery, or do they have a self-satisfied grin like they're pleased with a joke they've just told?

Then there is the identity of the person. Is it someone you know, and if so, whom? Is it your child who is screaming and red-faced, or your boss, or your spouse? And if it's a stranger, what kind of stranger: a little kid lost in the subway, or a strange man in a dark alley? Context matters enormously here, including location, time of day, and events that have immediately preceded the facial expression. Are you on a raft floating down a river in the jungle, or in Times Square on New Year's Eve? Has there just been an earthquake, or has a president just been elected? And your personal history, role in society, and cultural upbringing matter too: The same facial expression might be interpreted differently depending on whether you're American, Indian, Japanese, high-status or low-status, born in the 1950s or in the 2000s. Finally, of course, how you *react* to the facial expression requires additional stages of processing, with massive degrees of context-dependence there too. While some cases might not lead to making an inference about the person's underlying intentions, most probably do, and such computations—does he want to hit me, or does he need my help?—combined with information about the present context and one's own goals or attitudes vis-à-vis the person are likely to play a role in most decisions about how to react in response to the facial expression.

This kind of situation—heavily contextual, involving lots of interaction among different kinds of information, including many levels of experience and the unfolding dynamics of the social interaction—is generally regarded as the exact opposite of the kind of system that can be handled by modules. On a dual-systems view, then, what you'd want to invoke is general-purpose cognition. But does that really

make the most sense here? Or might it be just as plausible that the flexibility and context-sensitivity we see comes from the contingent interaction of many specialized systems, each contributing their computational labor to an emergent, yet designed, outcome?

Imagine you're walking down the hall at work and you see an object coming toward you. Early stages of visual processing compute the boundaries of the object and fill in color and depth information. Shape and motion cues are processed by animacy detectors, which tag the object as a living, animate thing, and object categorization mechanisms categorize it as a person. The blob sitting atop the person's shoulders satisfies the input conditions of the face detection system, which tags it as a face. This face tag then serves as a signal to face identification systems to feed information about the details of the person's facial features and their configuration to its search engine, which combs through its identity templates looking for a hit, probably using a massively parallel search. Within a second, a match is found: It's that guy from down the hall who, frankly, annoys the hell out of you.

At the same time as the facial identity search engine is churning away, the facial expression system has identified key points on the face such as eyes, eyebrows, mouth, and other identifying points in the facial musculature, and uses these to make an inference about what kind of facial expression is being produced. This includes assigning the expression to a basic emotion category such as fear, anger, happiness, or worry, but additional inferences are made about subtle variations on these basic emotion categories. For example, we might imagine that the person is generating an overt display of pleasure at seeing you, but rather than extreme pleasure, he is expressing a polite yet wooden sort of pleasure designed to prevent you from actually stopping and greeting him. This kind of subtlety is clearly something that (most) humans can detect.

These separate inferences—of the person's identity, the context of the encounter, and the nature of the person's facial expression—must now somehow be combined and submitted to several further inferential processes. For example, by combining these separate pieces of information, you might make an inference about the person's underlying mental state: While they aren't angry with you (good), they also don't particularly want to stop and talk to you (also good). This is just fine with you. It allows you to decide, in an instant, that it's fine to just nod at the person, politely smile in return, and keep walking—a course of action that some other search engine has pulled up from a menu of possible culturally appropriate actions, learned from experience as a member of your culture, as a member of this particular workplace, and as a member of this specific relationship.

Now, it might be that the cognitive division of labor here is not exactly as I've hypothesized. It could be that the computational problems are not carved exactly

at the joints I've proposed, though there is evidence for every one of these processes occurring. Regardless of the exact empirical details, several observations about the nature of the cognitive division of labor are in order.

First, it's clear that even though specialized processes are occurring in which different bits of information are being processed by distinct domain-specific mechanisms, the outputs of their computations must be able to *interact*. This is inconsistent with a pipeline model of cognition in which information, once it enters a domain-specific pathway, is in an encapsulated funnel that sequesters it away, never again to interact with the rest of the system. It's much more consistent with something like cloud computing—massively parallel, where lots of computations are occurring at once with an emergent result, as in Selfridge's pandemonium model. There are many conceptual and formal models of how such interactions might occur. These include parallel distributed processing models (Rumelhart et al., 1986–87), global workspace models (Baars, 1997), and multi-modular models such as Anderson's ACT-R framework (Anderson, 1996) and Peter Carruthers's model of an interactively modular mind (Carruthers, 2006). I have proposed the metaphor of a bulletin board, where representations are visible to multiple computational systems, each of which can add their processing to the representation (Barrett, 2005b). No matter how you conceptualize it, in order for the products of the different computations to be any use at all, there must be some stage or stages at which they are combined in a flexible way to produce higher-level inferences (e.g., about the person's intentions) and decisions (e.g., smile, nod, and keep walking). And one way in which specialized mechanisms might be designed to interact is through tags or informational signatures that each process adds to the growing representational object, such as "it's a person" and "it's *that* guy" and "he's smiling politely" and "he's not angry," which in turn can serve as inputs to other systems, such as intentional inference systems and systems for integrating cues to come to a behavioral decision.

Rather than occurring in a strictly ordered domino-like fashion that would be the same every time, interactive processes can be mixed and matched in different ways, and contextual effects can happen at many points in the processing. For example, it's widely held that the probability of recognizing someone's face can depend on context and not just on the person's facial features: You might be much more likely to recognize the guy from down the hall when you're at work, for example, than if you pass him in an airport on the other side of the world. This means that a supposedly automatic process like face recognition can be affected by contextual cues (Kerr & Winograd, 1982). And this could well be not a bug, but a design feature: using context to aid a search by making slightly more accessible face templates that you're likely to see in the current context. In fact, such contextual effects are common in cognition. They are often called priming effects and occur in diverse processes ranging

from interpreting word meanings to moral judgments (Tulving & Schacter, 1990). Information is brought to the fore that is likely to be useful in a particular context, making it more likely to produce a match in the time window where it's needed—a form of inductive bet. If this is true in the case of face recognition, it suggests that face templates are stored with contextual tags to facilitate searches.

Moreover, unlike a strict processing chain, a more interactive system in which different systems collaborate to generate inferences means that some steps can be skipped—some of the collaborating mechanisms can come up empty-handed—and inferences can still be made. For example, suppose that the person coming down the hall were a stranger. Your facial identity search engine would search in vain, finding no match. Does this mean that the facial expression interpretation system would be unable to produce an inference? Of course not. That system can still operate, producing an inference of a polite smile. The *meaning* of the smile, or the meaning you infer, might be different when expressed by a stranger versus the annoying guy; this is a matter of how the cues or tags generated by the various inference systems interact at the stages of intentional inference and decision-making. An ability to produce inferences from partial information could also be a feature, not a bug. In a strictly linear assembly line, if the sausage-stuffer seizes up, the whole assembly line freezes; in a more parallel, distributed system, one or more computations can be absent, and the system can still deliver an output. The ability of systems to operate even when some components fail is sometimes called *robustness*, and it's a property that brains and other biological systems, unlike manmade systems, have in abundance—probably not a coincidence (Kitano, 2004; Sporns et al., 2004).

---

What I have depicted here is of course just a sketch of what the real cognitive processes underlying a given social interaction might be like. Many of these sub-processes, such as emotion recognition and face recognition, are the focus of entire subfields of psychology and neuroscience. My point is to illustrate that the idea of flexible processes emerging from an orchestrated community of specialized components is actually quite biologically and neurologically plausible. It's at least as plausible as the idea of a rigid layer of ancient inflexible modules on the bottom with some more recent, general-purpose flexibility sauce poured on top. Indeed, the idea of a community of specialists collaborating in a designedly emergent way is an old one in cognitive science, dating back at least to Marvin Minksy's metaphor of a "society of mind" (Minsky, 1988). And the idea is consistent with much of contemporary neuroscience, which is increasingly providing evidence that the brain is locally modular, but globally massively interconnected, relying on coordinated, parallel activity among many

subsystems (Bullmore & Sporns, 2009). But is there more? Might there be something on top of all the modules supervising them?

To many, the answer is yes: On top of whatever specialized subsystems the mind might contain, and certainly distinct from them, is *consciousness*. Remember that several of Fodor's modularity criteria have to do with lack of interactivity with consciousness. Automaticity, for example, means that the operation of modules can't be consciously controlled, and encapsulation means that we can't consciously access what's going on inside them. This certainly suggests that whatever consciousness is, it's *not* the product of modules. On the other hand, the modular portrait of mental activity that I sketched above described lots of conscious mental activity, from the awareness of a person's identity, to the meaning of their expression, to a conscious choice about how to react. Is there a conflict between consciousness and the idea of a democracy of modules with none particularly in charge?

Maybe or maybe not. On the one hand, the distinction between conscious and unconscious mental processes is a widespread one, in all likelihood derived partly from our own subjective experience of our minds and partly from historical traditions and folk models (Freud, for example). It's also to some degree enshrined in dualistic models of the mind such as dual-systems theory, which dichotomizes mechanisms in terms of conscious versus unconscious control. These models suggest that whatever consciousness is, it's distinct from and not the product of batteries of evolved specializations.

On the other hand, the various emergence-based models of mental computation I described above—from pandemonium, to Baars's global workspace theory, to Minsky's society of mind—were all developed with an eye toward explaining conscious awareness and the sense of will. Importantly, they attempt to explain consciousness as emerging from an underlying set of interacting mechanisms, without postulating any magical extra bits. Whether they will be successful in doing so, of course, is a question we can't yet answer; it's hard to imagine a more contentious area of psychology than the study of consciousness. However, there's no question that many people view consciousness, and especially free will, as the pinnacle of mental activity that an evolutionary perspective—especially one involving massively parallel interaction of specialized mechanisms—can't explain. Before concluding this chapter, I'd like to take a brief detour to examine consciousness and ask: *Could* many aspects of consciousness, and perhaps all of it, be explained by the interaction of many specialized processes?

There is a long literature on consciousness that I won't attempt to summarize here, other than to say that the term, like innateness, has multiple meanings (Block, 1995; Block et al., 1997; Chalmers, 2010). Sometimes it means awareness, as in: I'm conscious of my heartbeat right now. Sometimes it means wakefulness or the ability to

experience, as when contrasted with unconsciousness. Sometimes it refers to volition or will, as in consciously choosing to do something. In all of these cases, the question is not whether some property is represented (somewhere) in the mind, but whether or not the representer is *aware* of that representation. Of course, the question of who or what the representer is—the "self"—is the million-dollar question. There is much psychological research on the self and the subjective sense of self. Many philosophers and psychologists suggest that it might be a kind of illusion (Kurzban, 2010; Wegner, 2002). What's clear, though, is that people can certainly *report* things such as what they are aware of, what they chose to do, and so on. In this sense, the phenomenon of consciousness is real. Philosophers use the term "phenomenology" to refer to those things that we consciously experience. And it's abundantly clear that there is lots of information in our minds that we're not aware of—most of it in fact—and a small amount of which we are at any given time. Why?

From an evolutionary point of view and from an empirical perspective, there are certain questions about consciousness that are more easily tackled than others. Philosopher David Chalmers has suggested that questions of what consciousness is, ontologically, are the "hard problems" (e.g., how consciousness is caused by physical stuff). Questions such as what information does or doesn't enter consciousness are the "easy problems" (Chalmers, 2010). The latter kinds of problems, to some degree, can be studied with ordinary scientific methods, and indeed have been studied since the early days of psychology by pioneers like William James. For example, you can simply put people in certain situations and ask them what they are aware of, as your doctor might do in a hearing test by asking you to raise your hand when you hear a sound. Or you can measure their reactions, such as whether they turn their head when you make a noise, or whether some feature of a visual display causes them to look. Of course, the latter technique has the problem that some reactions might not be consciously generated, and the former has the problem that not all conscious experience might be reportable. But at least in principle, these are starting points, and the bread and butter of contemporary experimental psychology.

From an evolutionary point of view, the key questions about consciousness are the same as for everything else we've looked at so far. What is its function, if any? What are its design features? How has it evolved? What I'd argue is that while consciousness certainly has an air of mystery about it, these questions are not particularly more difficult to answer in the case of consciousness than they are for other aspects of our psychology, as long as we frame our hypotheses carefully and don't attempt to bite off more than we can chew.

There is a popular view in philosophy that consciousness is an "epiphenomenon." On this view, consciousness is not, in fact, causal: We may be aware of things after the fact, but that awareness *itself* plays no causal role in shaping behavior (Kim, 1998).

While this could be true, one thing is certain: If it is, then consciousness has no evolved function, at least in the sense that we've defined function in this book. If there is no effect on fitness, there is no role of natural selection in shaping it, and therefore no evolved function. If that's true, then the theory and methods we're developing here will have nothing to say about it. Leaving that possibility open, then, let's consider the more interesting possibility that it *is* causal: that stuff enters consciousness for a reason, and that reason has to do with fitness. Then the questions become: Why do some things enter consciousness and some don't? What special role does conscious information play in shaping our behavior that unconscious information doesn't? And why does only a tiny amount of what's going on in our brains ever become conscious?

We have several hints to work with. The most obvious, as I mentioned, is that only a small amount of what's going on in our brains ever becomes conscious. By asking *what* that information is, we can get an inkling of why. A classic example is language parsing. As you read this sentence you are *aware* of its meaning: Some representation of its meaning enters your conscious awareness. But as we've seen, the chain of processing necessary to build (extract, infer) that meaning from patterns of ink on a page or dots on a screen is probably long and complex—and you're (mostly) not aware of *that*. You're not aware that this processing has occurred or of any of the rules or information in that processing chain (this is, indeed, what Fodor meant by encapsulation). Chomsky pointed this out too: All adult speakers of English, for example, know how to produce the future perfect tense ("I will have fallen asleep by the time you get home"), but most are not explicitly (consciously) aware of the rules used to form that construction. In fact, we have seen that this is normal for cognition in general. Semantic colorizers, for example, paint the inferred meanings of stimuli onto representations of them, such as identity onto a face; we are aware of the colors so produced, but not the mechanisms that do the painting.

There seems to be a clear, intuitive answer to the question of why we're only aware of, for example, the meaning of a sentence and not the rules used to parse it: It's the meaning we're after. Inferring the meaning of an utterance is, after all, the function of parsing. The rest is like scrap paper that we use to do a math problem: Once we have the answer, we throw the paper away. And indeed, that's probably part of the answer for why most of the processing in our mind is unconscious: It is the outputs of the process that natural selection has designed the system to produce, and those outputs therefore should play a special role in cognition.

This still doesn't answer our million-dollar question, however: Why *consciousness*? That is, why does this information need to be consciously represented to play a special role in behavior—and what is that special role? While we don't yet know for sure, it's likely that not *all* outputs of evolved computational processes are made

conscious. There are likely to be evolved procedures in your brain that produce outputs, such as updating your attitudes with regard to certain beliefs, adjusting your food preferences, and so on, of which you're not conscious. So why is there an elite club of representations to which only some gain access?

Again, we don't yet know the answer, but there is a family of conjectures in philosophy, psychology, and neuroscience that circle around the same idea: Consciousness represents a "bottleneck," a narrow information channel or workspace in which representations are weighed to reach a decision about how to allocate attention and control the body (Baars, 1997; Carruthers, 2006; Crick & Koch, 2003; Dehaene & Naccache, 2001; Pashler, 1984). The reason this bottleneck is narrow is that it contains only the most highly distilled outputs of mechanisms delivering fitness-relevant information and options in order to be able to decide rapidly and efficiently. And the reason that there is only one of them is that there is only one body to control: only one set of eyes whose attention to direct, only one set of lips to produce utterances, only one set of limbs with which to act.

On this view, what we're consciously aware of is the product of an underlying, unconscious pandemonium: an assembly of demons shouting, in Selfridge's metaphor, for control of attention and the body. Not only is this proposal consistent with the multi-modular view of the mind I have presented here, it makes sense of several empirical phenomena regarding consciousness and leads to some new predictions as well.

If the mind is a kind of neural democracy, then what we'd expect is a host of mechanisms and processes, operating simultaneously and in parallel, evaluating lots of stuff about the world and our moment-to-moment representations of it. Each of these processes would be designed to generate some outputs in the service of fitness: a feeling of hunger calibrated to represent the current nutritional state of the system, awareness of potential dangers in the immediate surroundings, rumination on a difficult decision problem at work, and such. Many of these processes can go on at an unconscious level, and only at a certain point do they become urgent enough that they require some bodily *action*: turning the head to look at a potential source of danger, picking up the phone to call the office, opening the fridge to look for food. If these processes are designed as an emergently orchestrated pandemonium, a neural democracy, then what we expect is that they will compete to insert the products of their computations into the consciousness bottleneck, as a function of how relatively important those outputs are at any given moment for controlling the body in the service of fitness. In Selfridge's terms, they will shout louder the more fitness-urgent their computations become, and the dynamics of competing neural activity create a kind of voting procedure that, when tipped sufficiently in one direction, will determine the outcome: awareness and, ultimately, a decision. Thus, the growing hunger

that you weren't really aware of suddenly pushes its way into consciousness in a way you can't ignore, or the system churning away at your problem in the office suddenly reaches a solution, creating a "eureka" sensation and dumping it into the consciousness bottleneck in a way that captures your attention and spurs you to pick up the phone.[5]

Subjectively, if you're like me, this captures at least something of what the stream of consciousness, and even the push and pull of conscious deciding, feels like. But there are also a variety of theories and suggestive evidence in psychology and neuroscience to support this model. The "neural democracy" idea is captured in neural network models that have different computations competing to attract attention, producing conscious experience (Crick & Koch, 2003; Koch & Ullman, 1987). The idea that unconscious computations place neural tags on their outputs to determine how those computations will be weighed in conscious decision-making is reflected in neuroscientist Antonio Damasio's concept of "somatic markers"—tags that make behavioral options feel a certain way: appealing, heavy, or effortful (Damasio, 1999). And there is evidence that people who lose the qualia, or feelings, associated with decisions become poor decision-makers (Bechara et al., 2000).

What this model predicts is that evolved mechanisms will be designed, at least in part, to interact as members of a neural democracy. In biology, it's recognized that an important factor influencing the design of multicellular organisms is that the different parts of those organisms must, in a sense, cooperate for the greater good in order for the whole body to survive (Maynard Smith & Szathmáry, 1995). Similarly, the diverse processes of mind must cooperate in running the body: Sometimes eating takes precedence, sometimes sleeping, sometimes reproduction, sometimes running from danger. In order to do this, each process must be designed to only take over the body when it's fitness-best to do so and step down when other processes should be given priority. Therefore, psychological processes should have design features—the equivalent of volume knobs on their outputs—that are calibrated to their role in the democracy (by natural selection, reaction norms, experience, interaction) and

[5] There are analogies to this kind of process elsewhere in biology. For example, there is the phenomenon of "quorum sensing," in which bacteria emit chemical signals that are used to regulate gene expression as a function of population-wide computations; quorum sensing can involve thresholds or tipping points that turn expression on or off in a switch-like manner (Miller & Bassler, 2001). It is also reminiscent of democracy-like decision-making in bees, the so-called wisdom of the hive (Seeley, 2009). Interestingly, Seeley calls this the "social physiology" of the colony, indicating the interplay between cognition and bodily processes—also, undoubtedly, an important link in consciousness as when we are aware of hunger, pain, the positions of our body parts, the effort of muscular exertion, and so on.

that are adjusted according to immediate circumstance. The literature on tradeoffs in decision-making suggests that something like this is going on—though of course, most of it hasn't yet been framed in the pandemonium sense I'm suggesting here (Payne et al., 1988).

If this model or something like it is correct, then it has implications for dualist views of mind. Perhaps the mind isn't composed of two layers, specialized and general purpose, but is instead more like a neural democracy. Nobody is really in charge; instead, multiple processes collaborate in shaping behavior, and what feels like "me" making a decision is really a kind of neural voting process emerging by design from a jostling assembly of demons. This doesn't mean, of course, that there couldn't be something extra needed to explain how the feeling of consciousness actually arises out of neural stuff—Chalmer's "hard problem." Nor does it mean that there couldn't be special mechanisms associated with conscious processes: for example, psychologist Alan Baddeley's proposal of a "visuospatial sketchpad," an "episodic buffer," and a "phonological loop" as components of the workspace of consciousness (Baddeley, 2003). But it would suggest that there doesn't need to be anything *but* specialized components interacting to produce all of cognition, all the way up.

Neural democracy models suggest that a two-layer model of mind, with a general-purpose homunculus sitting on top, is not mandated by first principles, and possibly not even by currently available data. Just as plausible is the emergence of consciousness from a confederation of specialized parts designed to collaborate in the control of the body and attention. They also suggest that thinking about the evolved functions of consciousness—at least in the form of hypotheses to be tested—might shed new light on an area of psychology that has long resisted empirical progress, and where various forms of dualism and Cartesian thinking have remained alive and well.

---

While the idea of a mind composed of parts is among the most contentious in contemporary psychology, I hope I have persuaded you that a biologically realistic version of the massive modularity thesis is not only plausible, but likely to be true. There *are* multiple processes in the mind with distinct functional features, and these interact to produce the whole of cognition. However, mental mechanisms aren't like Lego blocks in multiple senses. They aren't rigid and inflexible, either in the sense of being frozen eternally and innately in some ancestral past or in the sense of being unable to act contingently with other systems and to modify their behavior based on circumstance. Indeed, there is no reason why brain processes—modules—might not be able

to alter themselves dynamically in the course of processing, much as stomachs alter themselves dynamically in the course of digestion (e.g., via changes in gene expression, "contextual morphs"). They don't all have a standard format, like the buttons on Lego blocks that allow them to be snapped in and out of place. And indeed, they probably mostly can't be just snapped in and out of the mind without consequence. There is widespread debate in neuroscience about whether we should expect to see clean dissociations or a pure loss of an ability while leaving everything else intact when a mental mechanism is damaged. There is every reason to think (and evidence to suggest) that the answer is: a little bit, but almost never entirely, because stuff is interconnected and complicated (Shallice, 1988). However, what's crucial to retain about the idea of mind parts—whether we choose to use the word modules or the perhaps less loaded concept of a mental mechanism or mental adaptation—is the idea of specialized diversity. The brain is not all one undifferentiated blob; it contains different systems that do different things, even though it's massively interconnected, interdependent, interactive, and dynamic.

Beyond those basics—which, I'd argue, we know enough right now to see are probably true, and we ought to stop arguing about them—are the interesting details about specific mental mechanisms, their design features, and organization. This is the bread and butter of psychology and neuroscience, and as critics of evolutionary psychology like to point out, lots of it can be done without thinking about evolution at all. What I've tried to stress is the likely usefulness, indeed necessity, of thinking about mental adaptations *as* adaptations—if, of course, we drop the baggage of folk models of two-layer minds. If we do, I suggest, we can develop a model of mental architecture that consists of adaptations all the way up, from perception to consciousness, with no magical all-purpose ingredients needed.

To do so, we will need to think about how the various parts of the mind have evolved via descent with modification, and how they have been shaped to interact with each other as an orchestrated whole. This will involve thinking about mind parts not as newer bits that appear de novo, fully formed, and snap into older bits. The human mind is not, for example, a chimpanzee mind with a few extra modules or abilities added, just as human faces aren't chimp faces with human bits added on. Instead, human and chimp brains are evolutionary relatives of each other, each evolved through processes of descent with modification from ancestral brains. This doesn't mean that new parts can't evolve; it just means that they must evolve organically, through processes of differentiation and divergence from ancestral structures, just as new tissues or limbs or organs evolve. Here, we enter the territory of evolutionary developmental biology, or evo-devo, and the study of organisms as evolving ensembles of coordinated parts (Carroll, 2008). Just like the rest of the body and its various structures, the brain is a complex, hierarchically

organized organ: It is an adaptation, with adaptations nested within it, and adaptations nested within those, all designed to interact together to generate behavior. In order to understand the whole, we'll once again need to step away from outdated models of the mind to confront the biological reality of how complex assemblages of specialized parts evolve.

# 12

## WHOLES

Complexity is hard. By their very natures, complex systems bedevil attempts to think about causation in the ways that we are used to in everyday life. While there are vigorous debates about the nature of human causal cognition in psychology, what seems undeniable is that human minds have a propensity to want to distill complex causal interactions into what statisticians call simple effects: the effect of one variable on another, holding everything else constant. Even worse, there is a tendency to want to look for effects that are linear, in which the effect of one variable on another is either to constantly increase or decrease it, rather than a more complex function that might increase or decrease nonlinearly depending on the value of the variable. Add in the many other blind spots for thinking about causation in a brain like ours—designed for purposes like predicting the behavior of projectiles and people, but not understanding the evolution of itself—and the situation just gets worse.

When it comes to thinking about minds, this is a problem. Minds are made out of parts that interact, creating enormous nonlinearities in cognition. The parts do not instantiate only simple two-dimensional mapping functions. They reorganize themselves during development, and the whole system changes dynamically in the blink of an eye at the timescale of neural processing. Add in change over evolutionary time, in which change in one part of the system is virtually certain to have cascading effects on other parts of the system, it would seem as if our simple, linear, intuitive models of causation stand no chance.

These facts about human minds as objects of study, and the weaknesses of human minds as observers of those objects, are undeniable. But confusions about the implications of complexity have led to a series of non sequiturs in the psychological literature. For example, some have claimed that the dynamic, context-sensitive, and interactive nature of the mind means that it can't be modular (Gibbs & Van Orden, 2010). Others claim that because modules develop over ontogenetic time, they are not the result of natural selection, but rather the dynamic process of development (Karmiloff-Smith, 1992; Buller & Hardcastle, 2000). And some claim that because development is an emergent, interactive process, natural selection can't specify

phenotypic outcomes (Lickliter & Honeycutt, 2003). These claims seem to reflect a widespread view that there are two fundamentally different kinds of explanation for brain organization: one that explains brain organization as emerging dynamically from interactions at various timescales, and another that explains brain organization as evolving because of selection acting on phenotypes over evolutionary time.

If I've convinced you of nothing else in this book, I hope that I've convinced you that these are not, in fact, mutually exclusive explanations for the organization of the mind. On the contrary, they are going to have to be reconciled. It is both the case that the brain is a massively interactive dynamic system in which the interactions create the phenotypes that we see, such as behaviors, *and* that the way natural selection works is via a feedback loop between these phenotypes and the developmental systems that give rise to them. This is not backward causation: The events of the moment, the vicissitudes of emergent phenotypes encountering the environment, determine whether the reproductive resources that built those phenotypes get transmitted to the next generation.

This messes with our intuitions. For whatever reason, we'd like to separate the emergent products of developmental happenstance from what gets transmitted from parent to offspring, and yet the reason those things get transmitted is precisely *because* of developmental happenstance. Every behavioral event, every thought, every decision is a fleeting, emergent, never-again-to-be-repeated outcome of many interacting factors, and so it has always been: not just my decision about whether or not to eat this bagel right now or what words to type on this page, but also, in the past, Zog's decision about whether to walk down to the stream for water now or later, his offhand comment about the latest dust-up between Zig and Zug, and a nearly infinite number of such micro-decisions stretched across time and space and populations. None of these was ever the result of a single, innate, reflex-like module mapping linearly directly from cue to behavior, but was always the result of many interacting causal factors on many timescales. Still, if brain evolution is to have occurred, there must have been a causal pathway between the consequences of these decisions and the material stuff that emergently caused them. All evolution by natural selection involves retaining or discarding aspects of developmental design based on their interactive properties, and therefore all adaptations, at the level of phenotypes, are the result of selected emergence.

For us, this has two important implications. First, it means that we're going to have to grapple with the interactive, emergent nature of wholes head-on: whole genomes, whole developmental systems, whole bodies, whole brains—even, if the arguments in chapters 9 and 10 are right, whole societies. Second, it means that we're going to have to think about why those wholes have the properties they do in evolutionary, adaptationist terms: as the result of histories of selection in ancestral environments.

To properly understand how cognition works at the level of the whole brain, we're going to need to think about multiple scales of causation in space and time: the scale of real-time interaction between brain mechanisms to produce moment-to-moment cognition, the scale of developmental interaction that dynamically shapes phenotypes, the scale of population-level interaction that structures social and cultural environments, and the scale of evolutionary feedback in which alterations in one part of an organism change the adaptive landscape for the rest. Adding to the interactive complexity, of course, is the fact that these multiple scales are not independent at all: The reason that evolutionary changes push organismic design up or down fitness hills is because of effects those changes have on developmental and population-level interactions, and therefore on moment-to-moment cognition.

What we want to know, ultimately, is how the mind is organized, and why: What is the stuff it contains—both in terms of the developmental resources that build working brains and in terms of the brains that they build—and how did it get here? Sometimes, there is a temptation to think about human minds as a special case and to try to reverse engineer their particular properties as sui generis phenomena, but as we'll see, that's a mistake. It's not that human minds don't have some special features, but those features arose through a process of descent with modification from stuff that is far, far older than our split with chimpanzees. Therefore, we'll have to consider the whole enchilada: evolutionary history all the way back.

As we'll explore in greater depth in the next chapter, this means we have to resist the temptation to partition off just what's new or special in humans, since nothing in evolution is totally new, and even to the extent that we can talk about autapomorphies—uniquely derived features of human minds—we can only understand them in terms of the assemblage of other mechanisms within which they evolved. To borrow a term from anthropologist Gregory Bateson, we'll need to think about the entire "ecology of mind," the mind as a system of components interacting internally and externally, rather than considering the parts entirely in isolation (Bateson, 2000).

It also means that in thinking about the evolution of mind as a complex assemblage of parts, history matters: What happens at time A affects what happens at time B. Thus, we'll need to think about path dependence on a variety of timescales: neural, developmental, cultural, and evolutionary. At each of these scales, thinking about path dependence matters for understanding design, because the design features of mental mechanisms depend heavily on how they arose and how they were shaped, which in turn depends on what role they play in the ecology of mind. Deciding how to react to a person's comment depends on who that person is, which depends on face recognition, which depends on object processing. Adult abilities, like understanding speech and recognizing faces, depend on earlier developmental events and experiences. And later-evolving capacities, such as human-specific capacities for culture,

our elaborate theory of mind, and the capacity for spoken language depend on older capacities that enabled the evolution of these newer ones, such as adaptations for social living, simpler forms of mindreading, and preexisting forms of referential communication.

In this chapter, we'll turn from the question of the brain's individual components to the question of how these evolve as an interactive, functional whole. This won't involve ignoring the parts, because those parts—mental adaptations—are our focus of interest. Instead, we'll ask how evolution of the whole informs our understanding of the mind's individual components and how they are designed to develop and operate as members of a functionally designed collective.

---

When it comes to path dependence on the scale of information processing, it's important to be precise about what we mean. While it's logically true that there must be computational path dependence in that the neural events at time $t + 1$ must depend, causally, on events at time $t$, that doesn't mean, as I stressed in the last chapter, that we should always think about cognition as occurring via rigid, predetermined pipelines. In other words, computational path dependence doesn't necessarily imply fixed *pathways* of information.

If you think about other examples of ordinary causation, there is nothing so odd about this. For example, baking a cake exhibits path dependence in that steps come one after the other and the end result depends on the sequence of steps you take and the outcome of each, as well as the order in which you did them. But that doesn't mean that you couldn't, at any point, do something different and have the result come out differently in return. Path dependence is simply a consequence of the role of time's arrow in causation, and doesn't imply that causation is locked on rails.

Similarly, as I illustrated in the last chapter, ordinary cognitive outcomes like greeting someone in the hallway depend on a series of information-processing events, and if these events happen differently, the outcome can be different. Moreover, there is not necessarily just a single pathway through the system that results in adaptive outcomes. For example, as we saw, you might greet a stranger in a perfectly acceptable way without recognizing him, and yet greet him in an equally acceptable but different way if he were someone you knew. The same goes for the other contingencies we mentioned, such as the effects of different facial expressions, relationships, and immediate context, which can cause information to be processed differently than if those contingencies had been different.

We've seen a variety of models and metaphors for thinking about collaborative, distributed information processing: specialized mechanisms each attacking some

part of an information-processing problem and collaborating via division of labor to produce a solution. Let's consider what kinds of design features we might expect to see in such a system, and how new parts might evolve against a background of older ones.

Imagine a central pool of information, like a bulletin board, that can be examined by many specialized mechanisms all at once: In other words, their inputs are all connected to it neurally. Imagine that each of these mechanisms has input criteria such that when it detects a bit of information it can process, it does so, and then reposts the products of its processing back on the bulletin board for other mechanisms to work on (meaning, minimally, that the results of the processing add information to the bulletin board in some way). Let's say, for example, that this bulletin board initially contains all of the outputs of the initial stages of object parsing, up to the point where objects have been sorted into three-dimensional wholes, with shape and color information identified. If the eyes are directed toward a bear, then, this bulletin board contains a perceptually parsed representation of the bear with all of the perceptual cues that the rest of the brain might use, but with none of these tagged or interpreted for meaning (i.e., the bear has not yet been identified *as* a bear, its face has not yet been identified *as* a face, its snarl has not yet been identified *as* a snarl, and none of this has been fed into decision-making systems). Similarly, any object toward which attention was directed, from a tomato to a person, could be represented on this bulletin board as an object file with a collection of parsed perceptual cues.

Consider how specialized systems might digest the meaning of the perceptual representations on this bulletin board via a cognitive division of labor. We can imagine a battery of the sorts of devices I've described in previous chapters wired to this bulletin board, each processing stimuli that meet its input criteria. Object recognition systems, for example, might tag the object as a bear or a person based on shape criteria (this could be done hierarchically via a pandemonium-type process). Face recognition systems would identify the face and tag it as such. In the case of a face, distinct mechanisms would attempt to identify the individual and the facial expression. Motion-processing systems might attempt to parse and interpret the person's or the bear's actions. In each case, the appropriate mechanisms would operate contingently according to their input criteria, processing rules, and the inductive bets they instantiate, adding information to the basic perceptual representation if and when they were able. The result would be, in the case of the snarling bear, a perceptual representation of the bear with the additional information added that it's a bear, that's it's snarling, that it is therefore angry, and that it's approaching you. In the case of the annoying colleague, the result would be the perceptual representation of the colleague with information about the individual's identity, facial expression, and

approach behavior added. This representation, with various semantic tags inferred and added, is now ready to be sent to decision-making systems to interpret and compare with other information—including information about one's own state (fearful, confident), past experience (bears are dangerous, this guy is a jerk), and immediate context (if I run I might make it to that tree, my office door is only two steps away)—in order to reach a decision.

I've called this model the enzymatic model of cognition because it resembles the enzymatic systems that populate our cells, in which many specialized devices collaborate to carry out the order-generating biochemical work that keeps us alive (Barrett, 2005b). Enzymes perform the computations that make our cells run: they are large protein molecules which, by virtue of how their amino acid subunits are folded and arranged, catalyze chemical reactions, taking chemical inputs, processing them, and turning them into transformed outputs. Interestingly, while enzymes can be and often are arranged into physical assembly lines where one enzyme passes its outputs to another enzyme and so on up the chain, they need not be. An enzyme can float around in the cytoplasm, doing nothing until it contacts a substrate that satisfies its input conditions—meaning a chemical that bonds to its active site—upon which it processes the substrate, or substrates, catalyzing a reaction that transforms them into something else.

This is just a metaphor, of course. In this metaphor, cognition is like catalysis. What modules do is to digest information: Each device attacks some part of the whole problem and does its little part, the resulting whole of which is the extraction of meaning from the stimulus, interpretation of that meaning in light of the current situation, and eventually action. This can involve parallel processing by many enzymatic devices at once, as well as serial or stage-like processing chains where results of parallel processes are integrated and passed along to different processes, as well as top-down or horizontal control of processing by mechanisms elsewhere in the system. For example, visual inference of object identity is computed in stages, beginning with massively parallel feature processing starting in primary visual cortex and proceeding via the ventral "what" stream through the occipitotemporal and inferotemporal cortices, where complex objects such as faces are identified by integrating the smaller, hierarchical chunks processed earlier in the stream (Haxby et al., 2001; Riesenhuber & Poggio, 2002; Spiridon & Kanwisher, 2002; Ullman, 2007). Top-down information from the prefrontal cortex is also involved, reflecting the role of voluntary attention to certain features of the stimulus and voluntary retrieval of information from memory (Ganis et al., 2007). Crucially, not each of the processes that collaborate to recognize complex objects is the *same* process, even though the final result is the product of emergent interaction between specialized sub-processes. Moreover, object recognition mechanisms interact with different areas of the brain that compute and

respond to the significance of object identity for the organism, such as the orbitofrontal cortex and the amygdala, to generate a decision about how to react to it based on emotional coloration of the stimulus and other features (Adolphs et al., 1998; Bechara et al., 2000). For such collaborative processing to occur, there must be neural hubs, bulletin boards, or workspaces—the equivalent of cytoplasmic spaces in the enzyme metaphor—where information from many parallel processes comes together and is integrated to produce outcomes like inferences, decisions, and behavior. Indeed, there are probably many such workspaces at many levels of cognition in which enzymatic sub-processes collaborate to interpret complex stimuli. Consistent with this view, many brain networks are characterized by a hub-and-spoke design, in which diverse neural structures communicate through central nodes (Bullmore & Sporns, 2009; Sporns et al., 2004).

A second feature of this kind of model is processing cascades: chains of processing in which the outputs of one device alter how other devices will, or might, process the information. For example, a device that processes facial expressions might only be activated by information that contains a face tag, identifying a perceptual representation as a representation of a face and, perhaps, tagging certain relevant features of it (eyes, nose, mouth). In this case, a prior device dedicated to identifying faces would have to have processed the face and tagged it as such, and this tag would be part of the input conditions of the facial expression interpreting device. Crucially, the face detector would not be the same device responsible for facial expression analysis, nor would the facial expression interpreter be burdened with having to identify which things in the input are faces and which are not (Adolphs et al., 1994). Similarly, an imitation system might require a representation of a desired goal state in order to operate, but the mechanisms that generate motor plans and compare the result with the intended goal do not *themselves* generate the inference of the goal state; they use it as input. This is the whole point of cognitive division of labor: designed collaboration between devices.

What we know empirically about how the brain works suggests that something like processing cascades are quite likely to exist. We know that different aspects of information are processed in different areas of the brain, and presumably many of these processes rely on processing done by earlier systems. Crucially, however, the idea of workspaces and enzymes suggests that while processing cascades *can* be pipeline-like or wired in a chain, they need not be. Indeed, brain processes appear to be a complex mix of parallel and serial processes, involving lots of dynamical interplay between systems. Just like enzymatic cascades, which can have both parallel and serial processing elements, brain processes appear to exploit interaction between systems to produce complexly organized cognitive outcomes, such as an uttered sentence.

If processing cascades of this kind exist, and if they communicate in flexibly contingent ways, this has design implications; interactions can't be a willy-nilly mix of all possibilities, or outcomes would be random. Interacting systems must be designed, through natural selection and developmentally adjusting reaction norms, to be able to interact. Because this is a form of communication, it implies the neural equivalent of conventions. For example, if mechanism B decides whether and how to process information contingently on what mechanism A did to it, then there must be conventions of communication between A and B. B's design must make inductive bets about A's design, and vice-versa.

Suppose mechanism A is a face detector that scans object representations for features that are diagnostic of faces, and tags them as faces when those representations meet some threshold criteria. Suppose B is a mechanism that scans input for the face tag added by A, and when it encounters a representation with such a tag, it executes a search of a facial identity database, looking for a match of the current face to a known individual in the database. Such a division of labor certainly makes sense, but it requires a tagging convention: B must have design features that recognize the face tag added by A *as* a face tag.

You might argue that mechanism A needn't be designed to make any inductive bets about B, since B's processing is downstream of A. Indeed, it's quite possible that mechanism B, the facial identity mechanism, actually evolved later than A, the facial detection mechanism. We can imagine a species that only has A, the ability to detect *that* another member of its species is present, but not B, the ability to detect *who* it is (i.e., to remember and discriminate between individual members of the species on the basis of their facial features). If so, it might be the case that A's features predated those of B, and therefore it makes more sense to say that B evolved to make inductive bets about A, but not vice-versa.

There is something to this idea. Evolutionary biologists agree that there can be an asymmetry between newer and older mechanisms in which older mechanisms are more entrenched—harder to change, for reasons of path dependence that we'll get to below—than newer ones, which adapt to them (Gould, 1977; Wimsatt & Schank, 2004).[1] The evolutionary conservation of the same or similar developmental regulatory genes across diverse taxa is an example (Knoll & Carroll, 1999; Prud'homme et al., 2007). However, we mustn't conclude from this that psychological mechanisms

[1] This kind of process, in which evolution is more likely to tweak later events in development than earlier ones, partly accounts for the impression of early embryologists that "ontogeny recapitulates phylogeny," because it causes features of early development to be more phylogenetically conserved than later ones—though again, this is a generalization, not a strict law (Gould, 1977).

only adapt on their input end: Every mechanism sends its outputs somewhere, and those outputs can influence fitness, so they are subject to selection too. And it's not the case that just because a mechanism is evolutionarily old, it's frozen to future change. The appearance of a more recently evolved mechanism B can, in turn, alter selection on older mechanism A. Thinking about such interactions is the key to understanding evolutionary path dependence in the realm of cooperating devices.

---

What does it mean for a new brain mechanism to evolve? As we've seen, in evolution, the word "new" must be used with caution. The fact of reproduction means that anything new must in some sense be a variant of something old. When a new mutation arises in an organism's genome, for example, whatever effects it has on the phenotype aren't the result of that mutation alone. Its effects are projected into the phenotype via the developmental system that turns DNA sequences into organismic design, and that developmental system is of course a mosaic of elements both ancient and recent (Carroll, 2008).

That said, if we take a many-dimensional view of phenotypes, there is a sense in which we can define a new phenotypic variant quite precisely: A new variant occurs when a change in development—and this can be a change in DNA, epigenetic machinery, or even the environment, as long as it can be transmitted to offspring—produces a phenotype whose design extends into an area of phenotype space that was not visited before, in at least one dimension. To take a somewhat absurd example, we might imagine a species in which the tallest individual who had ever lived was six feet tall. If an individual appeared who was one inch taller than six feet, that variant would be novel in that its design hadn't yet been tested by selection: That tiny new area of design space hadn't been visited yet.

What's nice about this—nice from the point of view of cumulative evolution—is that new designs get to borrow heavily from old designs. Continuing with the example of increases in height, imagine that our species is an ancestor of modern-day giraffes, and the six-foot-one individual is able to reach leaves or fruit that its shorter conspecifics aren't able to reach. It certainly gets an incremental fitness boost from its added height. But what delivers that boost isn't *just* the added height: It's the legs, the neck, the lips, the tongue, and the visual system in combination with the height that deliver the extra food. Moreover, while we can imagine a single genetic mutation producing the extra height—and we do know that such height-increasing and decreasing genes exist (Sutter et al., 2007)—that mutation isn't by itself coding for a whole new chunk of neck, in the sense of providing the genetic information that produces an extra inch of bone, an extra inch of muscle, an extra inch of blood vessels, skin, fur,

nerves, and so on. That stuff is all provided for by the rest of the developmental system, including the rest of the organism's genome. The change is produced by the way the new mutation *interacts* with the developmental system.

Now, this might all seem obvious, but the implications for the origins of evolutionary innovations in the brain are apparently not obvious at all to critics of modularity. Indeed, it's common to think that when a new part or module appears, it must occur in exactly the way I caricatured above, as requiring new genetic code for *all* of the module's design, akin to an extra inch of neck requiring new genes for everything in that extra inch. This mistaken view is in part behind "not enough genes" critiques of massive modularity (Buller, 2005; Ehrlich, 2000; Tomasello, 1999). These critiques hold that every new module requires a big hunk of genes to code for it, and so differences in modules between species must be linearly related to differences in their genomes. If the brains of humans and chimps are 40% different in modular organization, for example, then their genomes should be 40% different. To these critics, the relatively modest genetic differences between humans and chimps must mean that differences between us can't be because of substantial differences in number or organization of modules.

As we're becoming increasingly aware through work in evo-devo, much evolutionary change involves rearranging existing developmental resources in new ways by alterations in gene expression (Carroll et al., 2005; Carroll, 2008). Thus, it's not just "one new mutation, one new trait." Large chunks of new design, in the sense I've defined it above, can be generated from relatively modest alterations in developmental programs. That's not to say that *some* genetic change isn't required. But as the giraffe neck example shows, the magnitude of the genetic change is not necessarily correlated with the size or complexity of the new chunk of phenotypic real estate you get. This is because, in general, the vast majority of the design in any new structure is not "coded for" by the mutations that give rise to that new structure. Perhaps the best known examples of this in evo-devo come from the study of limbs and the genetic regulatory machinery that causes them to appear in various parts of the organism at various times during development. Fiddling with those genes can, for example, produce entirely new limbs—even though the vast majority of the design features of those limbs aren't new at all and rely on many, many genes other than the one that's been changed (Carroll et al., 2005; Shubin, 2008).

Modifications in gene expression can result in limbs that are new in at least two senses. First (Sense 1), changes in expression can result in duplication of existing structures: a phenomenon sometimes known as *serial homology* or *paralogy* (Fitch, 1970; Hall, 1995, 2003; Koonin, 2005). In organisms with multiple limbs, such as insects, crustaceans, and humans, the different limbs of the organism (and even some parts not used for locomotion, like the feeding appendages of insects) are

actually evolutionary copies, or homologs, of each other. And this leads to the second sense (Sense 2) in which they are new: Modifications in gene expression can modify the individual features of the limbs, leading to differences between hands, feet, wings, fins, and mouthparts. This is functional specialization, and diversification, through descent with modification (Carroll et al., 2005; Hall, 2003; Ruvinsky & Gibson-Brown, 2000; Shubin, 2008). Maynard Smith and Szathmáry (1995) call this process "duplication and divergence," where "duplication" leads to novelty in Sense 1, and "divergence" leads to novelty in Sense 2.[2]

A variety of evolutionary biologists, psychologists, and neuroscientists have proposed that duplication and divergence of phenotypic structures is likely to be an important process in the evolution of new brain parts (Allman, 2000; Barrett, 2012; Clark, 2010; Kaas, 1984, 1989, Krubitzer & Huffman, 2000; Marcus, 2004, 2006; Griffiths, 1997; Matthen, 2007; Striedter, 2005). In this scenario, new brain tissue appears via mutation and/or alteration of gene expression, which right from the start inherits most or all of the design features of the brain tissue it evolved from (duplication). This is analogous to the inheritance of design of the extra inch of giraffe neck. Once this new tissue appears, it is then subject to further selection, which can select for features that gradually diverge in some ways from those features of the tissue from which it arose (divergence). In this way, new modules can appear that *already* exhibit substantial design, due to their homology with older ones. They are new, in Sense 1, at the moment they are born (and may indeed carry out novel functions right away, as we'll see below). They become new in Sense 2 if and when their design features are modified by natural selection, as a function of their role in the mental ecology into which they are born.

The conceptual taxonomy of homologies in evolutionary biology has grown quite complex in order to take into account when and how homologies—and in particular, homologous genes—originate in the evolutionary process. Just as phenotypic traits such as limbs originate through duplication and divergence, so do genes (Ohno, 1970). Geneticists have developed terminology to describe various forms of homology based on when and how new genes originate via duplication, and I have proposed

[2] I am not claiming that all novel design originates through duplication and divergence—far from it. Moreover, it's important to distinguish between duplication and divergence of phenotypes, and duplication and divergence of genes. The latter used to be considered a contender for the major driver of evolutionary change (Ohno, 1970). Plenty of cases of gene duplication exist, such as duplication and divergence of opsins, which we'll see below. But newer work in evo-devo suggests that much evolutionary change can occur without duplication of the underlying molecules. For example, many regulatory genes are highly conserved, and it is the way they act on different tissues that is modified (Carroll, 2008). This is still, of course, descent with modification.

borrowing this terminology to describe the origin of new brain structures via duplication and divergence (Barrett, 2012; Fitch, 1970; Koonin, 2005). *Orthologs* are good old-fashioned homologs: traits in different species that trace their ancestry back to the same trait in the common ancestor of those species. They have not been duplicated *within* the species, but duplicate copies of the same trait are present in two species after a speciation event. For example, chimpanzee and human amygdalas are orthologs, just as are the genes that code for chimpanzee and human hemoglobin proteins. In contrast, *paralogs* are serial homologies: multiple traits within a species that descend from a common ancestral trait somewhere back in the lineage (Fitch, 1970; Hall, 1995). To make things more complicated, we can distinguish inparalogs (duplicated traits that have arisen after a speciation event, i.e., serial homologs that are novel, or derived, in the descendents of that event) from outparalogs (serial homologies that were present in the ancestors of a current lineage, perhaps from very long ago) (Koonin, 2005).

This terminology was invented to describe processes of gene duplication, and we must be careful to realize that there is not a one-to-one mapping between genes and adaptations. The appearance of paralogous genes can produce serial homologies in the phenotype, though it doesn't have to, and serially homologous traits or characters needn't be, and probably usually aren't, caused by corresponding paralogous genes (Young & Wagner, 2011). However, if brain mechanisms evolve through descent with modification from ancestral structures, then this terminology is also potentially useful for thinking about how, when, and why new brain mechanisms evolve from old. For example, theory of mind mechanisms in humans and monkeys, if they descend from a common ancestral mechanism in earlier primates, are orthologs: descended from the same ancestral adaptation, but possibly modified in different lineages. Some aspects of our theory of mind mechanisms may be new in the sense of having been modified relative to other taxa, but old in the sense of being homologs. On the other hand, we might have developed new theory of mind mechanisms in Sense 1, above, through duplication and divergence; for example, if it's true that tracking *false* beliefs requires an entirely new mechanism, it might have originated through some sort of duplication and divergence event, and thus would be a paralog (and, if unique to humans, an inparalog). Still, despite its novelty in both senses I used above, it would be homologous to mechanisms present in other species.

Evolutionary biologists agree, and work in evolutionary developmental biology is increasingly confirming, that descent with modification is the primary mode of evolutionary change. This means that new traits leverage (and are constrained by) features of the ancient traits from which they evolve. It also means that novelty within lineages blooms like the branches of a tree from the roots of ancestral design, creating hierarchically organized design features (Barrett, 2012). Duplication and

divergence of both genes and phenotypes is likely to play a role in this diversification of traits within organisms.

In the case of genes, a good example is the opsins, which we first encountered in chapter 1. In humans, three distinct opsins are responsible for our three-color vision, and these are paralogs, having evolved from a common ancestral opsin (Dulai et al., 1999). The three are tuned to be maximally sensitive to red, green, and blue wavelengths in the visible light spectrum, respectively. Thus, there is a common ancestral design for changing conformation in response to light energy—the first step in transducing light into a neural signal—but divergence in the domain of light wavelengths each is sensitive to. Thus, opsins have design features in common, yet each is also sub-specialized. Specialization, in this case, is hierarchical: specialization (wavelength domain) nested within specialization (the ability to change conformation in response to light and to trigger a signaling cascade).

There is currently a debate in evolutionary genetics about the processes that result in divergence of design following duplication of a gene. An older model, proposed by Ohno (1970), held that when gene duplication occurred—as an accident of gene replication—the original copy could continue to carry out its evolved function, leaving the second copy free to vary. In other words, alterations in design of the second copy wouldn't disrupt the evolved function being carried out by the first copy of the gene, so the second copy could change either via drift or selection to carry out some new, similar, or slightly different function (or in some cases, the second copy could be silenced if not net beneficial). Another model, sometimes called "subfunctionalization," holds that the two copies could diverge via a process similar to what is known in evolutionary biology as adaptive radiation or niche differentiation to carve up the functional space into distinct regions or components, with one copy diverging to cover one part of the space, and the other copy covering the other (He & Zhang, 2005; Lynch & Force, 2000; Rastogi & Liberles, 2005). Something like this seems to have occurred in the evolution of opsins, which diverged to handle different parts of the light frequency spectrum.

As the case of limb evolution shows, duplication and divergence of phenotypic traits can occur as well. While this can occur through duplication and divergence of underlying genes, it needn't, or at least not entirely. This is because, among other things, the ontogenetic differentiation of phenotypic structures can occur through gene regulation. Indeed, as evolutionary developmental biologist Sean Carroll points out, "toolkit" genes, which are like master switches that orchestrate development of morphological features like limbs—the *Hox* genes being the best-known example—are rarely duplicated (Carroll, 2008). This makes sense given their central function as orchestrators of functional sub-differentiation of tissues, and the path dependence of evolution. Duplication and divergence occur, in essence, on top of these genes, which

remain stable. Thus, the *same* gene, in combination with contextually activated regulatory genes, results in building an arm here and a leg there. This is, indeed, common in evolution: Anciently evolved building blocks remain highly conserved, and they in some ways play the same role across specialized parts of an organism (e.g., "build structure of homolog type X"), but in other ways play contextually differentiated roles, building homolog type $X_1$ in one place (e.g., a leg) and type $X_2$ in another (e.g., an arm).

Processes of hierarchical specialization are likely to be important in brain evolution as well (Barrett, 2012; Kaas, 1984, 1989; Krubitzer & Huffman, 2000; Marcus, 2004, 2006; Striedter, 2005). If brains increase in complexity over evolutionary time in the same way that everything else in organisms does—through increasing functional sub-differentiation of specialized parts—this has implications for thinking about mind design. In particular, it means that many commonly held views about modules are wrong. Each module is *not* a sui generis phenomenon, independently evolved, with its own unique suite of genes. Each new module does not start at the bottom of the design hill: Because new structures inherit the design of the older structures they evolve from, new modules begin far up the hill already. And if modules evolve via processes of hierarchical specialization from common ancestors, some design features will be widely shared across modules, others within more narrow subgroups of modules, and some may be specific to individual modules.

This is quite different from the idea that specialized adaptations are frozen in the past, and evolutionary change occurs by adding general-purpose plasticity, in the form of more cortex, on top of them. Instead, as cortex expands, it inherits an enormous amount of specialized design, including design that results from interactions with other brain regions. This inherited design is then the starting point for subsequent evolutionary modification. If this turns out to be the major mode of evolutionary change in brains, it has important implications for thinking about how minds evolve not as collections of independent and unrelated parts, but as coordinated, coevolving wholes.

---

As the first law suggests, what it means to inherit and modify design depends entirely on the details of the design being inherited. How is information-processing design instantiated in neural tissue?

Some neuroscientists would tell you that it isn't. Or rather, they'd say that design isn't inherent in neurons or genes themselves, but emerges through general-purpose wiring rules of neural systems, interacting with their inputs. For example, neuroscientists Quartz and Sejnowski claim that "the developing cerebral cortex is largely

free of domain-specific structure," and "the representational properties of cortex are built by the nature of the problem domain confronting it" (Quartz & Sejnowski, 1997, p. 537).

As for many of the other claims we've examined in this book, there is a sense in which this is right, and a sense in which it's wrong. It's right in the various ways that we've talked about emergence accounts of development being right: Neural structures definitely *do* emerge as a result of interaction with the structure of their inputs, and therefore many of the details of the resulting structure cannot be explained without recourse to the structure of the input. But it's wrong in all the ways that we've talked about emergence as a false alternative to design: Emergence and natural selection are not mutually exclusive explanations for the phenotypes we see. In fact, natural selection can and must influence the design of the outcomes of neural development at many levels.

For example, there is the process that is sometimes called "modularization." Developmental psychologist Annette Karmiloff-Smith has pointed out that many skills can become modularized in the sense that over developmental time, areas of neural tissue become dedicated to handling these skills and increasingly specialized or tuned to just the skills in question (Karmiloff-Smith, 1992). This can include skills that have a relatively clear and plausible evolutionary history, such as the dedication of the facial fusiform area to the task of face recognition (Kanwisher & Yovel, 2006). But they can also include skills whose evolutionary history, if any, appears less clear, like bike riding, chess playing, and reading words (Dehaene, 2009). Based on this, Karmiloff-Smith makes a similar argument to Sejnowski and Quartz: Modularization is not the result of natural selection specifying modules, but results from a process of emergence.

While there's clearly something to this idea—bike riding, chess playing, and reading are all evolutionarily novel—treating this as an either/or issue, innate versus learned, seems too simple. Instead, this seems an ideal case for applying the concept of an open reaction norm and the type/token distinction. There is in fact a range of options for how specific modularization processes are to particular types of input. Just as all opsins share an opsin-general mechanism for light sensitivity but individual opsin types have their own specific wavelength thresholds, the design features of systems that modularize brain regions during development may be hierarchical (Barrett, 2012; Bassett et al., 2011). What I mean by this is that Karmiloff-Smith and Sejnowski and Quartz are right that there *are* general modularization procedures that are shared across many neural systems and that have design features that result in progressive specialization and tuning of modules based on experience—but there are also likely to be modularization features that have been tuned to different types of

modules in the brain as a function of where and when they develop, and what kinds of inputs get fed to them.

Here, we might think about adaptations for *module spawning*: the creation and shaping of modules, during development, via the interaction of informational inputs with developmental reaction norms (Barrett, 2012). This could include both procedures for shaping the properties of modules via inputs, and also for spawning *new* modules via a process akin to the developmental bifurcation of tissues into subtypes, or the spawning of new limbs. In the case of brain development, this is sometimes called *parcellation* (Johnson & Vecera, 1996). As elsewhere, one can imagine a range of possibilities for module spawning reaction norms, from extremely general ones to ones more narrowly tailored to specific kinds of inputs, brain regions, or developmental contexts. For example, one can imagine properties common to many neural systems (such as strengthening or weakening connections as a function of input statistics) that could lead to parcellation. Psychologist Robert Jacobs has used neural networks to model a bifurcation process that splits a network when it receives different kinds of input that are most efficiently handled via developmental subfunctionalization (Jacobs, 1997).

One could also imagine module spawning processes that are hierarchical in nature, with some modularization procedures shared widely and others more narrowly. Consistent with this possibility is evidence that specialization itself, as instantiated in developed brain phenotypes, often appears to be hierarchical. Neuroscientists Stanislas Dehaene and Laurent Cohen (2007) have pointed out that many cortical areas consist of modules within modules. Some areas of cortex, for example, are map-like, such as areas involved in object recognition: Shapes in the world can be mapped to brain areas that detect them in a quasi-topological manner (Tanaka, 2003). Dehaene and Cohen suggest that these cortical maps are organized hierarchically into macromaps, mesomaps, and micromaps. Structures that handle fine-grained aspects of inputs (micromaps) are nested within higher-level maps (meso and macromaps) that handle more global aggregations of features in pandemonium style. If so, there are presumably modularization procedures that create this hierarchical organization during development, using input to tune the maps but in a manner that has hierarchically organized modules as its developmental target.

In neuroscience, it's common to think of brain tissue, or at least the outer layer of the cortex, as just a general-purpose putty with the same properties everywhere. In principle, however, natural selection is not limited to a single set of modularization procedures common to all tissue. Consider, for example, the hierarchical nature of limb development driven by differential expression of genes in different tissues during development. First, large-scale differentiation into body segments occurs, and then these segments differentiate as a function of where in the body they are

and when they develop, with local interactions causing the differential expression of genes (Carroll et al., 2005).

Similarly, in the brain, there are good reasons to think that differentiation of tissues is hierarchically organized, controlled by regulatory systems, emergent in the sense of being driven by local interactions, *and* designed to produce specific types of modules, in specific places, tailored to handle particular types of information. It's difficult to study gene expression in the brain in vivo, in either real time or during development (in humans, at least). But there is growing evidence that brain development occurs via the differential regulation in genes at different times and places in the brain during development, mediated by factors such as microRNAs (Khaitovich et al., 2004; Krichevsky et al., 2003; Krubitzer & Huffman, 2000; Krubitzer & Kahn, 2003; Miyashita-Lin et al., 1999; Nakagawa et al., 1999; Namihira et al., 2008; Zeng, 2009).

There are at least three ways that natural selection can shape the specialization of brain parts during development, which I'll call where, when, and what-based differentiation. Where-based differentiation involves switching on or off regulatory machinery as a function of where in the brain the neurons in question are developing (Carroll, 2008; Hall, 2003; Khaitovich et al., 2004). We know that there is some sense in which all neurons are the same, in terms of basic features of neural functioning such as sodium channels and the like. However, that doesn't mean that neurons can't organize themselves differently, as a matter of design, in different tissues, or that different genes can't be turned on or off in neurons as a function of where they are, when they developed, or what information is passing through them (West et al., 2001). For example, neural organization differs in different brain regions, presumably in part because of differential gene expression as a function of location within the brain (Lein et al., 2007). And while brain anatomists believe that such region-based differentiation occurs at fairly large scales, this may partly be the result of looking at fairly large-scale properties of tissue that can be seen easily under a microscope, such as densities of connectivity. There is no reason that there couldn't be more fine-grained differences in regional wiring rules that correspond to more subtle aspects of structure—the neural equivalent of sensitivity to different light wavelengths—that aren't easily seen under a microscope. Such where-based differentiation processes would be analogous to those that cause arms and legs to develop differently based on their positions in the body, pinkies and thumbs and big and small toes to develop differently as a function of where they are on the limb, and so on down to even finer levels of spatial detail (Carroll et al., 2005).

When-based differentiation involves differential gene expression over time (Carroll, 2008; Hall, 2003; Wen et al., 1998). This is where developmental path dependence can enter the picture: A tissue can develop differently if developmental event X

has occurred than if that event hadn't happened. There are many known cases of when-based differentiation in development, and they tend to interact with "where" information: For example, the development of organs like breasts and testes and structures like teeth and hair depend on a combination of where and when. In the brain, as we will see below, this can include the dependence of development on inputs from adjacent neural tissues that have developed prior to the structure in question, thereby providing it with certain inputs needed for it to specialize. Again, while this *can* be merely emergent, it can also be the result of designed emergence, in which the emerging developmental events have already been experienced over many generations and tuned by selection.

Finally, there is what-based differentiation: specialization of neural tissues on the basis of what information is being fed into them during development. Presumably this is primarily what Quartz and Sejnowski have in mind when they say "the representational properties of cortex are built by the nature of the problem domain confronting it" (Quartz & Sejnowski, 1997, p. 537). For example, a neural region might get progressively turned into a specialized chess module because it's being fed information about chess and being tuned by neural rules such as Hebbian wiring rules (Gilbert et al., 2001). On the hierarchical design account I'm proposing, we would expect a variety of such what-based tuning mechanisms that would be common to and inherited by a wide range of brain structures. But on the hierarchical account, we'd also expect the possibility of *contingent* what-based procedures: procedures activated when a developing tissue region begins to experience particular types of input. For example, there may be informational signatures corresponding to motor activity that are different from the signatures of object representations, setting the stage for the evolution of different developmental procedures to be switched on or off when such signatures are encountered. And there may be many, many more such informational types that activate contingent developmental procedures. What I've been calling representational tags—which must have distinct neural signatures in order to be recognized as such by the other mechanisms that use them—could equally well serve as informational triggers to turn on or off developmental procedures in tissues that receive those tags in their neural inputs. It's common in neuroscience to assume that content-based differentiation of neural tissue is entirely the product of domain-specific inputs being fed to domain-general neural tuning processes. However, studies showing that tissue can be rewired to take on the properties of another specialized cortical area, such as auditory cortex, by feeding it the appropriate inputs are also consistent with the possibility of specialized reaction norms being activated by some informational signature in the input (Sur & Leamey, 2001).

Let us pause to consider some general features of this hierarchical model of brain specialization. First, it's biologically plausible. The origination of modules

via duplication and divergence, and the tuning of their design via modifications of the developmental regulatory machinery that gives rise to them are consistent with the general lessons of organismic development from the field of evolutionary developmental biology. They are also consistent with general principles of how brain evolution occurs via modification of ancestral designs (Krubitzer & Kahn, 2003; Striedter, 2005).

Second, this model of module evolution and development is quite different from conventional views about modules, which hold them to be entirely innate and require them to originate via entirely new chunks of phenotypic design and correspondingly large chunks of genome. It allows for the evolution of specialization without requiring that every single detail of phenotypic outcomes—or even most of them, for that matter—are specified in the genes. For example, it's easy to envision modularization processes specialized to produce types that, while highly variable, have certain features in common, from object representations (lions to sports cars), to practiced and stereotyped motor skills (–shooting arrows to riding bikes), to the grammars and lexicons of languages (Armenian to Zulu), to cultural norms and rules (handshakes to turn signals).

Recently, several authors have proposed what is called the "neural reuse" hypothesis: that older brain structures can become co-opted for new purposes (Anderson, 2010; Dehaene & Cohen, 2007). While this is sometimes seen as inconsistent with a "preformationist" view of modularity, it's certainly consistent with the model I'm proposing here. For example, as Stanislas Dehaene and colleagues have documented, exposure to a lifetime of reading written words appears to result in a modularized region of the object recognition area of the cortex, specialized for reading words, sometimes called the "Visual Word Form Area" (Dehaene, 2009; Dehaene & Cohen, 2007; McCandliss et al., 2003). This area exhibits many properties one would expect of a specialized object recognition area tuned specifically to words, including preferential activation upon exposure to words and a flexible, pandemonium-like process in which words are recognized equally well across a variety of fonts and orientations. However, we should be careful what we mean when we speak of an evolutionarily new area here. On the one hand, it's clear that written characters have only existed for a few thousand years. However, there are several lines of evidence suggesting that this is a spawned paralog of other object recognition areas, including spatial proximity to other object recognition areas (for faces, places, and tools) and design features shared with object recognition more generally, such as recognition procedures robust to orientation and partial occlusion (Dehaene, 2009). Moreover, the spawning of a novel module in this case may represent a case of felicitous novelty, in which the inputs—words—satisfy the inductive bets of object recognition systems, and not by chance. Psychologist Marc Changizi and colleagues provide intriguing evidence that

characters in the world's writing systems have evolved via processes of cultural evolution to mimic naturally occurring objects and therefore to be easily processed by human object recognition systems (Changizi & Shimojo, 2005; Dehaene, 2009). If so, this is a paradigm case of mind-world fit resulting from evolved reaction norms operating on culturally coevolved inputs.

Finally, this model suggests that blanket claims that the entire cortical region of the brain contains no domain-specific design features are somewhat premature. If specialization were instantiated via regulatory processes differentially activated based on local where, when, and what contexts, and if it relied heavily on input structure to guide development but in at least some locally contingent ways, how could we distinguish mere emergence from designed emergence based on the kinds of data we have now? For the most part, we couldn't—not, at least, with anatomical dissections, or fMRI data, or evidence from brain damage. Instead, we'll need to start looking explicitly for the patterns of brain differentiation we'd expect on this model, which would include differential gene expression in the brain as a function of when, where, and what is occurring in different tissues. In fact, we already know this is quite likely to be true based on work with other animals like mice and monkeys (Krubitzer & Huffman, 2000; Nakagawa et al., 1999; Wen et al., 1998), as well as more recent studies of humans and chimpanzees (Khaitovich et al., 2004). This means we're going to have to start entertaining more specific hypotheses about just *how* regional specialization is occurring during development, and these are going to have to look beyond merely the general properties of neurons.

---

How might we go about developing such hypotheses? The two major ingredients are path dependence and natural selection. We've now seen in abundance that while these are different elements in the toolbox of evolutionary causation and evolutionary explanation, they are not independent: It's not descent *and* modification, it's descent *with* modification.

Let us consider a case study, the evolution of a specialized adaptation for recognizing faces (Kanwisher & Yovel, 2006). On the hierarchical specialization view, it's plausible that this system evolved from more ancient object recognition systems via a process of descent with modification (Barrett, 2012). We can imagine an initial state in some ancestral animal that had the capacity to parse objects into object files and to create different categories of objects and discriminate between them. The creation and use of such categories would of course occur via learning, but it would have to be a form of learning that was specialized in certain ways. For example, it might involve module spawning procedures that created templates for assigning objects to

particular categories, where a new template—a modular structure with input criteria tuned to the category in question—was created under certain circumstances. A monkey, for instance, might spawn a token object recognition module specific to detecting leopards, given the right kinds of experience with leopards (including, for example, observing conspecifics exhibiting fear of them; Cheney & Seyfarth, 2007; Mineka et al., 1984).

Such spawning events would depend on the inductive bets of the module-spawning system, which could include things like whether or not the shapes of certain objects exhibited statistical clustering, and criteria for deciding whether a particular clustering was fitness-relevant enough to warrant generating a new neural template to detect it (Jacobs, 1997; Tenenbaum et al., 2011). Those criteria could in turn use feedback from other systems like emotion systems. For example, a predator avoidance system that relied on agency detection and certain kinds of motion cues (rapid approach) to decide to flee an object might activate a procedure in object parsing systems that looked for commonalities between objects that had chased you, and when enough commonalities were found, an object template for that class of predators would be spawned. Social cueing of fear could play a role as well.

This initial, ancestral object recognition system is of course already quite complex and the result of lots of prior evolution. Hopefully, you can imagine at least a rough picture of the kinds of steps it might have taken to get there. Importantly, such an ancestral state would have had much complex design already in place: things like object parsers, animate motion detectors, behavioral decision-making systems, and the capacity to spawn new modules on the basis of experience. In this scenario, the module spawners of interest would be downstream of object parsing, both neurally (in the flow of information through the brain) and developmentally (in the order of what gets constructed before what). During development, the module spawners would use the outputs of object parsers in combination with the other kinds of criteria mentioned above, which are also the result of outputs of other systems (e.g., agency detectors and action parsers) in order to cause a new, category-specific object recognition template to be built (a predator template). This template would be an inparalog of other object recognition templates.

The spawning of this new, inparalogous object recognition template for predators would have occurred by design, in the sense that the module-spawning procedure in question evolved to expect object representations as inputs and to spawn new modules based on the statistical properties (e.g., clustering and distinctness) of the inputs flowing in. This does not necessarily mean that a new, specially designed object spawning algorithm appeared de novo right at this place in developmental time and space in the brain. Module spawning might be part of a brain-wide neural toolbox that can be turned on at certain times, places, and contexts of development via the

activation of regulatory machinery present in all neurons. One might imagine that occasionally, module-spawning procedures are turned on or off at random via alterations either to gene sequences (mutations) or, even more likely, local modifications to epigenetic markers on DNA. Perhaps most of the time, such regulatory changes would not be fitness-enhancing, but when they were—and if they were inheritable—they would be retained. The turning on of module spawning just downstream of object parsers might be an example of such a fitness-enhancing change, and if so, it would be retained. At first, a module spawner downstream of object parsing might not have properties *specific* to spawning new object recognition modules—such as a module for predators—though once activated there, it could begin to be shaped by selection specifically for that purpose.

Now imagine that at some point in the evolutionary history of a species—possibly an ancestral species of monkey, or perhaps an even more distant ancestor—it became advantageous to pay attention specifically to faces as sources of information. At first, remember, this might not entail recognition of the faces of particular individuals, but just detecting that a face was there. We've seen, for example, that detecting eyes and the general direction of gaze including facial orientation can be quite useful even in simple, reflex-like systems like broken wing displays and tonic immobility. Initially, these might appear via any kind of mutation or neural variant that caused faces to be differentially reacted to in a fitness-enhancing way: Some initially very crude shared features of faces are lumped together by the object system's commonality-detecting algorithms, causing an alteration in behavior (e.g., approach, avoidance, inspection) that has a slight fitness effect. This would be the first step on a pathway up a fitness hill, initiating selection *for* the reliable spawning of a face detector during development based on some combination of where, when, and what criteria. This face detector might initially start as just an ordinary token of an object template spawned by the object template spawner, but once it started systematically getting face input, it could become increasingly specialized for faces. Geographically, this scenario makes sense given what we know about where face recognition is located in the brain, in the facial fusiform gyrus of the temporal lobes, just downstream of more general object detection mechanisms and adjacent to mechanisms specialized for recognizing other kinds of objects, such as locations (Kanwisher & Yovel, 2006).

The next step in our scenario, once a paralogous mechanism specifically processing faces has appeared via duplication, is the modification of this region to discriminate between *individual* faces, based on inductive bets tuned to faces in general (e.g., bets about similarities and differences in human facial configurations; Maurer et al., 2002). Such a system would also be selected to interact with other databases of information about individuals such as category membership (friend, foe, mom, mate) and individual relationship histories. However, note that selection for an ability to

distinguish between individuals—based on facial features or anything else—depends on certain other things being true about the social ecology of the species. Jellyfish, for example, might not get a fitness advantage from a novel mutation that allows them to recognize other individual jellyfish. Such selection depends on there being a fitness advantage to individualized dyadic relationships, be they friendly (e.g., dyadic alliances) or antagonistic (e.g., recognizing the individual who is alpha in a dominance hierarchy). And there are other elements of face processing as well, such as emotion processing, that probably involve at least partly distinct mechanisms from facial identity recognition and that require separate evolutionary benefits (Haxby et al., 2000).

While we don't yet know for sure, there is evidence that specialized face recognition abilities may be widespread among primates (Pascalis & Bachevalier, 1998). This is consistent with the social ecology of many primates, who are known to be able to recognize other individuals (by faces, voices, and other cues), and who get fitness benefits from discriminating between other individuals as friends, kin, and members of dominance hierarchies (Cheney & Seyfarth, 2007). Importantly, we can't understand the evolution of this specialized cognitive mechanism, face recognition, except in the context of evolutionary history (e.g., as a homolog of other object recognition abilities, appearing within a mental ecology that already parses objects) and in the context of a particular kind of social environment. Moreover, there are likely to be feedback dynamics once face recognition originates, allowing for more and more fine-grained personal relationships once individuals can tell each other apart.

Let's take stock of what this scenario suggests about the evolution of newer parts within older wholes. First, while I've picked the evolution of face detection and face recognition as an example, duplication-and-divergence processes could explain the evolution of specialized architecture elsewhere in the mind as well. New parts can appear via various forms of path-dependent inheritance of design, meaning that the evolution of new parts doesn't require whole-cloth, de novo appearance of giant chunks of new design, or what are sometimes called "hopeful monsters." And this doesn't necessarily mean that developmental designs must be inherited immediately adjacent to or downstream from where they originated, either spatially or temporally: The idea of module-spawning procedures being turned on or off in different parts of the brain, contingent on local events, shows this.

Second, this example shows something about the designed collaboration or designed emergence that we should expect to see in cognitive and developmental wholes. New systems evolve to make inductive bets about systems that are already in place, and these inductive bets evolve via a process of duplication and divergence. Initially, for example, module spawning machinery that appears downstream of object parsing might make no inductive bets *specific* to objects, but if it is reliably

activated there, it can start to be shaped by selection to assume or expect that its inputs will be parsed into objects. Similarly for parts specialized to detect faces or recognize individuals: Their inductive bets evolve because these systems continue to receive the same kinds of inputs over generations, providing a signal of variances and invariances that selection can operate on.

What this implies is the evolution of conventions of communication between modules. The concept of evolving communicative conventions is what makes sense, neurally and developmentally, of what I've called informational tags. A tag would originate as any neural signature that was consistently or statistically present in the output of an already existing system A that would be reliably present in the inputs of a newly evolving system B, and that provided a signal for natural selection to operate on in shaping the inductive bets of B. In some cases, the tag could be an entirely arbitrary feature of A's output that could simply be used by B to know *that* it is receiving output from A. In developmental time, this tag could be what turns on the regulatory machinery that begins to build a module of type B. In neural processing time, it could be what an enzymatic system uses to know that it's encountered an instance of a certain type of input. The path-dependent nature of this is that A doesn't have to "know" in advance that B is going to evolve, develop, or use its input, because B has evolved to exploit a signal that was already present before it was there, on any of these timescales.

However, it's important to remember that once some later-evolving system B is in place, future generations will inherit *both* system A and system B. This means that for A, B is now part of the whole in which it operates, and the presence of B can now exert selection on A, changing its fitness landscape. This is why bidirectional conventions of communication can evolve. Once B is in place and consuming the outputs of A, natural selection can then act on A to alter its outputs in ways that make B's operations fitness-better. In this sense, for example, object parsing systems can package object representations in ways that are *designed* for later systems to use—and this can be true even when some of those systems didn't exist when object parsers first evolved. Thus, while path dependence is critically important in brain evolution, it's also important to avoid the mistaken impression that phylogenetically older mechanisms are frozen in time, unable to be altered by later evolutionary events.

---

What is the role of genes in all this? Up until recently in biology, and certainly in the intuitive model of genes that most of us have, the role of genes has been thought to be essentially static. This is why it's easy to think about genes as "coding for" aspects of the phenotype, or as providing instructions to be read, or even as containing little

pictures of the phenotypic structures they are said to specify. But as we now know, genes are much more active than this, much more dynamic: They can be turned on and off, sometimes on timescales of minutes (Cai et al., 2008). This means that the contingent expression of genes and the activity of their products—including effects on other genes, as well as many other kinds of interactions within and across cells—can have causal effects at many spatial and temporal scales. We've seen that genes can have spatially and temporally contingent effects in brain development. But gene regulation might even play a role in neuronal activity on behavioral timescales, altering what neurons are doing in response to immediate circumstance.

I have mentioned several examples of gene regulation that change the activity of cells on relatively short timescales, such as the role of the lac operon in turning on and off the ability to digest lactose depending on its presence in the environment (Jacob & Monod, 1961) and the role of insulin in coordinating the behavior of many other genes in response to food (O'Brien & Granner, 1996). These gene regulatory responses are each a kind of computational system that responds to changes in internal and external environments. And like any computational system, each has inputs (e.g., the presence or absence of lactose), operations (upregulation or downregulation of genes), and outputs (digestion of lactose). Since Jacob and Monod's initial discovery of the lac operon, many other such genetic on-off switches have been discovered, including some that affect behavior and cognition. I mentioned, for example, genes activated by learning in mouse brains (Guzowski et al., 2001; Hall et al., 2000), and genes whose expression correlates with particular behaviors, such as nursing and foraging, in honeybees (Ben-Shahar et al., 2002; Toth & Robinson, 2007; Whitfield et al., 2003). Indeed, evidence suggests that differential expression of genes across brain tissues is common, though our understanding of how, when, and why these genes are turned on and off is still in its infancy (Khaitovich et al., 2004; Lein et al., 2007).

Currently, it's common to think of the computational properties of neurons as governed solely by a small set of responses common to all neurons that are contingent on just one thing: neurotransmitter pulses across the synapses between cells (Gallistel & King, 2009). On this view, the main way in which neurons compute is via transmitting these signals down the axon and across to another neuron, and altering their responsiveness to further stimulation as a function of the spatiotemporal combinations of firings they've received from the other neurons attached to them as inputs. In essence, then, the entire brain is made of one kind of neural Lego piece: a signal-transmitting cable with a volume knob and a single learning function for setting it, common to all the Lego pieces.

But what if neurons could signal to each other in other ways? And what if neurons could alter their algorithms for deciding when to fire and how to alter their firing patterns, not just as a function of the neural inputs they are receiving but also as

a function of where they sit in the brain's architecture, what kinds of information are passing through them at the moment, and what has passed through them in the past? In fact, what we now know about gene regulation and the local contingencies it permits—including cell-cell signaling within and across regions of tissue—opens a host of possibilities for how neurons might become "tuned" in domain-specific ways based on where, when, and what information (Jaenisch & Bird, 2003; West et al., 2001; Zeng, 2009). We don't yet know much about this, because the study of in vivo gene expression is new and difficult to do in functioning adult brains. But if neurons can differentiate into types and even regulate their activity from moment to moment as a function of local events of many kinds, this opens up room for functional complexity and specialization far beyond what might have seemed possible with a brain made solely of one kind of Lego block.

When molecules and the expression systems that produce them are highly conserved, as is the case with most biological signaling molecules, the result can be "dumb molecules, smart tissue": A single gene expression product can have different effects on different tissues. The molecule can in some sense be arbitrary in its form, an evolved signaling convention to which different tissues evolve their own responses (Torday & Rehan, 2012). This is known to be the case in body development, where high conserved gene regulatory systems like homeobox have different effects on different tissues, while sharing a common function in a global sense (i.e., orchestrating limb development) (Carroll et al., 2005). The same is true in the nervous system, whose basic components are highly conserved—many, in fact, predating the origin of neurons themselves (Zakon, 2012). Signals may be emitted by one tissue that then orchestrates the turning on and off of genes in other tissues. Dopamine, serotonin, oxytocin, and cortisol are all examples. However, just because these gene expression products are highly evolutionarily conserved does not mean that they can't play a role in diverse, specialized adaptations. Hox, for example, helps build limbs in fruit flies and humans, but the resulting limbs are quite different in form.

Interestingly, comparative work suggests that gene expression pathways influencing behavior may be highly conserved. For example, a particular type of gene product, cyclic AMP-dependent protein, plays a role in memory from insects to mammals, and another, cyclic GMP-protein kinase, plays a role in foraging behavior across diverse taxa of insects, as well as worms (Toth & Robinson, 2007). As elsewhere in biology, it is likely that evolution tinkers with these ancient, phylogenetically widespread gene expression systems to produce taxon-specific behavioral tuning. Vasopressin and oxytocin, for example, are two gene expression products that are thought to play a role in social behaviors like parental investment and maternal bonding, and these regulatory systems likely play a role in forms of human sociality that have both ancient and derived elements, such as cooperation (Keverne &

Curley, 2004; Kosfeld et al., 2005). And in addition to proteins, microRNAs are likely to play an important role in both nervous system development and functioning—a whole new control system that we are just beginning to understand (Zeng, 2009). In all likelihood then, brains aren't just one kind of Lego block arranged in different ways entirely as the result of firing patterns passing through them. In addition to the synaptic transmission properties of neurons, gene expression is likely to provide another, parallel "keyboard" that can be played, metaphorically speaking, to tune the properties of neurons and orchestrate their behavior.

---

The picture of brain evolution I've attempted to paint here is, I think, quite different from widely held caricatures of modularity and evolutionary psychological views of the mind more generally. I'm not advocating a preformationist view, but rather a designed emergence view. On this view, phenomena like domain specificity and specialization do not reside in little pictures in the genome. They emerge from interaction among the genome, its associated regulatory machinery, and external and internal environments. This is a constructivist view of development: Phenotypes are not simply unpacked from a box and plugged in; they are built during development. But contrary to some constructivist views, there is no reason why the only source of the domain specificity that we see in developed brain phenotypes need be the world itself, any more than the differentiation of different limbs or tissues during development is caused entirely by a domain-specific world interacting with a domain-general genome. Instead, evolutionary developmental biology is showing us that developmental systems have a variety of means for instantiating developmental contingency. Unless the principles of brain development are radically different—and much simpler—than the rest of development, there is no reason why brain development can't be selected to have domain-specific developmental procedures of many kinds, both broadly and narrowly scaled.

This does not mean, of course, that we currently understand what domain-specific processes structure the cortex, or even if there are any. It could be that there are none. But I'd wager on something less than full cortical equipotentiality. In addition to the theoretical reasons I've given here, there are many empirical ones too. For example, neuroscientists have known for many decades that the actual phenotype of the developed adult human cortex is organized into domains (Penfield & Rasmussen, 1950). Not only are different brain regions dedicated to different functions, those specializations are distributed across the surface of the brain in ways that are not only difficult to predict a priori just from the way the world is structured, but that are uncannily similar across individual humans. This does not mean, of course, *total invariance*

across individuals. But if you accept the arguments I've made in this book about variances and invariances, blueprint-like identicalness isn't a prerequisite for evolved specialization.

Domain-specific organization in the brain exists hierarchically at many scales, just as the hierarchical modularization view I've presented would lead us to expect. The evidence is so massive that there is no way of reviewing it all here (see, e.g., Cabeza & Nyberg, 2000; Gazzaniga, 2009; Mazziotta et al., 2001; Posner & Raichle, 1998; Price & Friston, 2005). Just for starters, if you open any neuroanatomy textbook, you'll see that the cortex—the outer layer of the brain—is divided broadly into lobes (occipital, temporal, parietal, frontal), and even at this global scale, there is functional specialization. Early visual processing, such as edge detection, occurs in the occipital lobe, for example, and complex object recognition of things like faces, hands, and bodies occurs in the temporal lobe. In bands across the top of the brain are the somatosensory and somatomotor cortices, first explored in detail by neuroscientist Wilder Penfield in the 1940s (Penfield & Jasper, 1954). The somatosensory cortex is responsible for registering bodily sensations, such as touch, and mapping them to the correct part of the body. The somatomotor cortex is responsible for orchestrating motor activity, such as reaching, grasping, and playing tennis.

The mere existence of separate sensory and motor regions suggests domain specificity, but there is also evidence that these regions exhibit substantial differences in design as well, in accordance with their distinct functions. Penfield and colleagues discovered that the somatosensory cortex is organized as a kind of map of the human body, known as the "somatosensory homunculus," which develops similarly in most humans. This map is arranged topographically but with some distortions: Most parts on the map are anatomically adjacent if they are adjacent on the body, but not all, and some parts (like the lips) are grossly distorted compared to the real thing, corresponding to the density of sense receptors in that region of the body. A map-like organization makes sense given the function of the somatosensory cortex. Interestingly, scientists were puzzled for a long time by the apparent non-map-like nature of the somatomotor cortex: Different bits of the body, like fingers, appeared to be represented at multiple places in the cortex haphazardly. As it turns out, the somatomotor cortex appears to be largely organized around actions, like reaching and grasping (Stepniewska et al., 2011). Because the same body part can be involved in different actions, the somatomotor cortex is not map-like, but rather like a set of routines corresponding to different functionally organized motor movements. This is consistent with the different functional specializations of these regions and suggests the possibility of functionally distinct reaction norms that build them, funneling experience into the construction of neural tissue in different, specialized ways.

The somatomotor and somatosensory regions are just two long-recognized examples of specialized, domain-specific regions in the cortex. There is evidence for many more as a result of new technologies of brain mapping. For example, evidence suggests specialized regions where language production and language comprehension develop, where mirror neurons occur, in the what/where pathways, regions involved in processing faces, places, bodies, mental states, and probably many more (Decety & Grèzes, 1999; Grossman et al., 2010; Hickok & Poeppel, 2007; Kanwisher, 2010; Price, 2000; Rizzolatti & Craighero, 2004; Saxe & Wexler, 2005; Ungerleider & Haxby, 1994). However, there is debate about how to characterize the precise functions of each of these regions. Even in cases where functional specialization is acknowledged, researchers sometimes disagree about what the region in question is specialized for. For example, we've seen debate about whether face recognition areas are evolutionarily specialized *for* face recognition or a wider category of objects (Gauthier & Nelson, 2001; Kanwisher, 2000). Another example is an area involved in speech production, sometimes called Broca's area. It used to be thought that this area had evolved specifically for speech, but now it is recognized that there are homologous areas in other primates that don't talk. How much it has been modified in our lineage specifically for speech, then, is a matter of debate (Friederici, 2009; Hickok & Poeppel, 2007). However, while there is debate about the exact nature of the specialization in each case, there is no doubt that these regions are specialized: Broca's area doesn't handle object perception, and the facial fusiform gyrus does not handle speech production.

This does not mean, of course, that there is not a substantial degree of plasticity in how and where these develop, nor does anything in the developmental account I've given suggest that it should be otherwise. We can't be certain yet that there is anything specialized in the developmental systems that build, for example, the somatosensory cortex that is different from the systems that build the visual cortex. But, given that these develop similarly across humans in terms of where they develop, when they develop, and what they process, what reason do we have to think that there *aren't* any specialized procedures involved?

A second aspect of the account I've given here that differs from conventional views of modularity is that it is based on interaction, and interaction by design, at all three of the causal scales we've been discussing: evolutionary, developmental, and neural. New modules evolve right from the get-go based on how they interact with other modules in the system. This means that their design features can only be understood in light of such interactions. Interaction is not just part of the function of evolved mechanisms, but in a sense all of it: Every single one of them is designed to get information from somewhere and to send it somewhere else. This is true developmentally and in real time. If development occurs via reaction norms that are

instantiated in gene regulatory systems, and if developmental systems make use of where, when, and what information in shaping phenotypes, then the emergence of specialized design during development is also the result of designed interaction. In real time, if modules evolved to interact via conventions of communication and to build complex representations through processing cascades, their operations can't be understood in isolation either. Indeed, evidence is mounting that virtually all brain activity involves massive interactions between regions. There is domain specificity in how information is processed in the brain, but it often involves domain-specific patterns of interaction, not isolated islands of activity that are solely responsible for a particular process or task (Roland & Zilles, 1998).This is very different than at least one cartoon version of modularity in which modules are isolated, autonomous, and inflexible, each operating without regard to what the others are doing and able to be snapped in and out like Lego blocks without affecting the operation of the rest of the system.

This view of modules means that interactions with the world can shape development, all the way up and all the way down, and yet it's still modular and specialized. In a species like ours that's sensitive to cultural inputs, this means that culture has the potential to shape developing brain phenotypes at all levels. This can extend from object recognition systems that allow us to recognize printed letters, to the way we categorize colors, to how we parse and remember events, to culturally novel but modularized skills like bike riding or car driving or videogame playing, to how we solve reasoning problems, think about time, or navigate space (Boroditsky, 2011; Dehaene, 2009; Levinson, 2003). While findings like these are typically taken as evidence against an evolutionary psychological view of a mind composed of specialized, evolved modules, I'm suggesting they shouldn't be, or at least shouldn't *necessarily* be. Whether data speak for or against evolutionary psychological hypotheses depends on having the proper type description of the kinds of modules and modular processes the brain has been selected to generate. Whether particular tokens are novel or non-novel, variable or invariant, isn't enough by itself to rule evolutionary design hypotheses in or out, absent hypotheses about types and how they get built.

The view I'm presenting here is also radically inconsistent with many current views of the relationships between genes and brain modules, which hold that each new module needs its own proprietary chunk of genomic real estate, and therefore that the human genome doesn't have enough genes to account for a modular mind (Ehrlich, 2000). Since that's not true for limbs, tissue types, or other modular anatomical parts, there's no reason to expect it to be true for brain parts. This isn't to say that new structures don't usually need at least *some* new genes, but the relationship among the number of parts, the complexity of parts, and size of the genome involved is not at all linear, contrary to what a "one module, one block of genes"

view might predict. What evo-devo shows us is that small genetic changes can lead to large phenotypic changes. This means that "not enough genes" arguments, which claim that there aren't enough genes in the genome to account for a massively modular brain and/or that there aren't enough genetic differences between humans and other species (e.g., chimps) to account for modular differences between them, don't hold water. Moreover, the regulatory module-spawning view proposed here is consistent with findings of large differences in gene expression in chimp and human brains, despite only modest sequence differences (Enard et al., 2002; Khaitovich et al., 2006). This is what you would expect if the basic building blocks for spawning modules are heavily conserved, but utilized partly differently in different species (though probably with greater degrees of overlap than divergence). This is not to say, of course, that sequence differences would not be required for specialized modifications to phylogenetically older module-building and brain- organization procedures. But the sequence changes required to produce substantial organizational and functional changes might be relatively modest compared to the changes that might be observed in regulatory activity.

Finally, the view of modular evolution I've presented here presents a major challenge for what I call "one reason" theories of human uniqueness: theories that attempt to isolate a single cause for the many differences we see between us and chimpanzees, or any other animal. If you take a complex developmental system with many hierarchically nested reaction norms controlled by a vast set of locally contingent, regulatory on-off switches and start tinkering with them—as well as adding new chunks of brain, including whole new banks of modules, via brain expansion—how do you identify just those aspects of human brains that are new compared to our evolutionary ancestors? And should you expect the answer to consist of one or two ingredients that you sprinkle on a chimp in order to get a human? Unfortunately for those seeking the holy grail of human uniqueness, the answer is likely to be a lot less simple than we'd like.

# 13

# Us

Humans are special—or so we like to think. Of course, there are many ways in which we're just like any other animal. A few trips to the doctor will convince you of this. And yet, there are ways in which humans appear to be remarkable, even among the astounding diversity of life. We've expanded from a rather small and unremarkable population of apes in Africa several million years ago to a worldwide population of several billion humans, altering the face of the globe in the process. The details of our everyday lives—some details, anyway—seem distinctly different from those of other apes. We live in houses, decorate ourselves, cook our food, sing and dance, talk—and in some places, drive around in cars, stare at glowing screens full of pictures and letters, and update our Facebook pages. To date, as far as I am aware, no other animal has been found that updates its own Facebook page, much less invents Facebook. What is the explanation for all this?

As you might be aware, human nature is currently one of the most contentious ideas in the social sciences. If you were not aware of this fact, you might be surprised. Most people take for granted that there is a human nature and might find disputes about its very existence to rank up there with other "academic" arguments. However, just because lots of people take something for granted doesn't mean that we should accept it as real. In the case of human nature, there are lots of intuitions about it, but what should a properly biological account of human nature look like—if any?

Debates about human nature involve several nested problems. First, some philosophers and biologists argue that the fuzzy and variable nature of species and populations means that treating any species as having a "nature" is a kind of category error—species just aren't that kind of thing (Hull, 1976; Mayr, 1982). The fact that humans share nearly 99% of our genes with our close relatives (depending on how one computes genetic overlap) is seen by some as damning to the idea of human nature (Ehrlich, 2000). Anthropologists have challenged the idea of human nature by pointing to the variability of human societies as evidence against the idea of universals, and by pointing to the fact that most of the data we use to make inferences about human nature come from Western societies (Henrich et al., 2010; Sahlins, 2008). As

we've seen, there is precious little agreement about what the basic components of human cognition might be—something that would seem necessary to come to an agreement about human nature. And there is the sticky problem of what, if anything, is *unique* to humans. Some people say that the answer is, in essence, nothing, except for the plasticity that comes from an expanded neocortex. Others have proposed a variety of candidates that might be the magic ingredient that makes us special, from language, to sociality, to the ability to use symbols (Deacon, 1997; Donald, 2001; Penn et al., 2008; Tomasello, 1999). And yet, there seems no consensus about what that magic ingredient might be, if any.

Here, briefly, I'd like to consider what the view of mental evolution I've presented implies for thinking about human nature. As you might expect, I'll argue that many of the ideas out there about human nature are not likely to be correct. And yet, I also think that *some* notion of human nature is indispensible. Indeed, I'd suggest that if most social scientists *really* didn't think there was such a thing as human nature, they'd probably quit their jobs and do something else. For many of us, we're in this because we want to understand *us.* For that to be a worthwhile endeavor, there's got to be something to understand. I think there is, but as for everything else we've examined, we've got to be careful. Whatever human nature is, it's a biological phenomenon, with all that implies. This means we have to be mindful of where we get our assumptions and how we apply them.

---

The very idea that species have natures is sometimes said to be "essentialist" (Mayr, 1982). Essentialism is a mode of construal in which things are treated as having some shared underlying "essence" that gives rise to surface features that can sometimes vary (Barrett, 2001; Bloom, 2004; Gelman, 2003). Another way of saying this is that essentialism lumps things together as "natural kinds." Examples of natural kinds that philosophers refer to are water, gold, and diamonds. These can take on different shapes, for example, but water is always water because of its molecular structure, as are gold and diamonds. Additionally, water can't change its essence without becoming something other than water; in this sense, essences are immutable.

Philosophers debate the ontology of natural kinds. Are there truly any natural kinds in the universe, and if so, are biological species one of them? Probably not, if certain philosophical criteria are stringently applied. Because populations change over time, for example, the "immutability" assumption of essentialism doesn't hold (Mayr, 1982). And yet to judge from newspaper accounts and television documentaries, some people think it makes sense to ask when the first human appeared. Paleoanthropologists feed into this illusion by placing some timeframe on the origin

of our genus, *Homo*, or by placing a date on the split between human and chimpanzee lineages. However, this is an artifact of the Linnaean classification scheme, which requires drawing category lines across continua in cases of speciation. There has indeed been an unbroken *lineage* of human ancestors leading to us, but there was no first human—it was all just continuous change.

What about a shared underlying essence, even if it does change gradually? Here again, the technical answer is probably no. All humans on earth do share a vast majority of their DNA. When measured in terms of percent sequence overlap (a dicey proposition, because not all DNA has the same functional significance), we're something like 99% similar. However, it's not nearly as easy to point to the DNA we *all* have in common—gene loci where everyone on earth is identical—because there are lots of rare polymorphisms. And even if you could, would you want to say that all and only those loci are what make us human? For example, if you found one person on earth with a mutation in a gene where everyone else is identical (something that happens, because all of us carry idiosyncratic mutations), would you say that they are no longer human because they are missing one of the critical genes? Or conversely, would you want to say that a gene in which a few people carry a rare polymorphism can't, by definition, be important in building traits that are characteristically human?

Probably not. And yet, the idea that these intellectual parlor games invalidate the idea of a human nature of some kind seems to make too much of philosophical definitions, since humans as a whole clearly present phenomena to explain. Indeed, there are those who on the one hand are skeptical of the idea of human nature, and yet on the other hand make statements like "humans are social" or "humans are a cultural species" or "language deeply structures human affairs." I agree with all of these statements, but I also think we couldn't make them unless there was something to refer to; otherwise, they'd be semantically empty.

We can agree, then, that essentialism—at least as I've defined it above—is an ontologically incorrect stance when applied to humans or to any other biological species. However, this doesn't mean there aren't reasons we tend to be essentialist about humans, other species, or even groups within a species, like sexes. For one, it's plausible that there are circumstances in which essentialism can be *inductively* useful, even if technically wrong (Barrett, 2001; Gelman, 2003). If true, this could explain why essentialist views of human nature—such that there was some flipping point at which ape turned into human—are so widespread and difficult to eradicate. But this doesn't mean, of course, that our scientific theories of human nature should be essentialist. How, then, *should* we think about human nature?

While we certainly need to jettison essentialism, this doesn't necessarily rule out other options. Biologists refer to things like "turtles," "orchids," "primates," and "*Homo sapiens*," usually without undue confusion. Even other biologists, the very people who reject essentialism, know what they are talking about. So there must be something in the world these terms refer to.

This doesn't mean that there isn't a debate in biology over how to define "species"—there is (Wheeler & Meier, 2000). One of the biggest problems is that there doesn't seem to be a definition that works well for all living things. For example, one of the most popular species concepts is the so-called biological species concept, which treats individuals as part of a common species if they are able to interbreed. This concept runs into problems for organisms that *don't* interbreed: asexual organisms. However, the biological species concept works reasonably well for some species, such as contemporary humans; it's reasonable to say that we are all members of the same species because any human male and female can mate to produce offspring. There are, of course, exceptions: sterility, fatal birth defects, and the like. There is endless philosophical handwringing about such cases, but if we accept that species are not "all-or-none," but rather probabilistic and cloudlike, they don't seem terribly worrisome. The point is that there are plenty of ways, biologically, to think of humans as a species while accepting the crucial biological fact of variation and that there are exceptions to every rule.[1]

That said, even among those who accept that all living humans are members of the same species—as nearly everyone does—the question of human universals remains contentious. For example, there are some who would like to define human nature as whatever is universal in humans, and would therefore deny that something can be a part of human nature if there is variation in it. There are some who insist that human nature is an illusion, a product of a Western worldview (Sahlins, 2008). And it's certainly true that people routinely mistake features of their own local culture for features of humans in general. For example, many Americans currently believe that marriage is "and has always been" between one man and one woman, an assumption that any anthropologist, not to mention the Bible, could tell you is incorrect. As anthropologist Joseph Henrich

[1] The ability to sequence ancient DNA from Neanderthals and Denisovans has reopened debates about where the "boundaries" of the human species lie (Reich et al., 2010). The very fact that we use distinct names like "Neanderthal," "Denisovan," and "*Homo sapiens*" implies that there are clear categories, but evidence suggests substantial exchange of DNA across these category boundaries. From a biological perspective, this is not particularly surprising, given the ability of these species to interbreed. But this is yet another aspect in which strictly categorical notions of human nature are challenged.

and colleagues put it in a recent article, much of the data we have about human psychology and behavior comes from societies that are "WEIRD": Western, Educated, Industrial, Rich, and Democratic (Henrich et al., 2010). Most psychology studies, from which conclusions about human nature are often drawn, are done on college undergraduates. And no matter how you slice it, it's a mistake to generalize from a non-representative sample to an entire population (Barrett et al., 2012; Sears, 1986).

While these empirical points are well taken, it's important to be clear that universality in the strict sense is not a necessary criterion for human nature. Certainly, as we've seen, universality of *phenotype* is not necessarily something we expect even in evolved adaptations; it depends on the shapes of the reaction norms that build them. It could be, for example, that no single detail of grammar is found in *all* the world's languages (Evans & Levinson, 2009). And yet we wouldn't want to conclude from this that there can't be adaptations for grammar acquisition. Similarly, while there is undoubtedly substantial moral variation across the globe, and it might be impossible to find a single moral rule that is identical everywhere, this doesn't mean that human morality can't be shaped by underlying adaptations, some of which could even be human-specific. Culture, language, mindreading, cooperation among non-kin—all are things that I, for one, would want to count as candidate aspects of human nature, even though the way these things play out in human phenotypes is kaleidoscopic, not monochromatic.

On the view I've been advocating here, both variation and lack thereof (if any) should be of interest to those studying humans, and both are important sources of evidence about design. Even on a probabilistic, population-minded, reaction-norm-based view of human nature, we eventually want answers about how variation is structured across humans. If we lay the measuring tape across the range of human phenotypes, some dimensions are clearly more variable than others, and what's more, the *reasons* for the variation or lack thereof should be of interest to us. If human phenotypes are the outcomes of evolved developmental systems being deployed across a range of conditions, then what we want to understand is how and why the diversity we see is produced. This is not contrary to an adaptationist approach to human nature; it is how such an approach should be done.

---

Yet another sticking point in debates about human nature is the question of what is unique or special to humans. Some people would like to equate human nature with human uniqueness, shaving off everything that's shared with other species and leaving just those things that are unique and derived in our lineage. On this view, the

remaining set of uniquely taxon-specific traits—autapomorphies—would constitute human nature.

As I suggested in the last chapter, I don't think this will do, for several reasons. One has to do with how we use the concept of human nature, or any species' nature for that matter, in our everyday thought and language. Many people would claim, for example, that a mother's love for her children is part of human nature—and yet, of course, maternal care of offspring is a trait that is found in most mammals, and not unique to humans at all. It's part of a lion's nature to chase prey, but not uniquely so. To err is human, but we can't take all the credit. There are many things about ourselves that are crucial to understanding who and what we are, but that aren't unique to our lineage.

Then there is the more technical problem with neatly separating derived traits from ancestral traits. Because everything (or nearly everything) in evolution evolves through descent with modification, all derived traits are in some ways versions or modifications of ancestral traits (Hall, 2003). This is not to say that there *aren't* derived changes in the human lineage that make us different from, for example, chimpanzees; there must be. Advances in genomics and epigenetics are allowing us, for example, to find differences in gene sequences and gene regulation in human and chimpanzee lineages. In some cases, we can even figure out (or hypothesize) why these changes occurred, and fancy statistical techniques allow us to examine which changes are the result of recent selection, as opposed to drift (Novembre & Di Rienzo, 2009). However, as the extra inch of giraffe neck example showed us, it's not always easy to partition the resulting change in design and phenotype into the old and the new. Moreover, even if we could, it's not clear that we'd want to do such a partitioning as part of our notion of human nature. Let's say, for example, that some genetic and epigenetic changes led to the expansion of frontal lobes in humans—a nearly certain bet. Would we then want to say that human nature is those mutations alone, but that all the cortical design features that are paralogously duplicated (and perhaps modified) in the resulting tissue—even though the tissue itself doesn't exist in chimps—are not? The resulting view of human nature would be bizarre to say the least, and probably not of much use in understanding human psychology or behavior.

As a side note, the strict uniqueness view of human nature would leave us with a strange way of thinking about sex differences. Sex differences are notoriously one of the most contentious topics in social science, and yet it seems under-recognized that many or most of the sex differences that are typically proposed in humans are homologous with those seen in other primates. Not only that: Most human sex differences, like the degree of body size dimorphism, are *smaller* than in other primates. This is thought to be because over the past few million years, human males have come to invest more in their offspring than the chimp-human common ancestor probably

did, leading to lower between-sex variation in investment in offspring and therefore reduced differences in sexual selection on males and females (Kaplan et al., 2000). A human uniqueness view would force you to say that sex differences aren't part of human nature because they are homologous with sex differences in other species—though it's unclear how you'd talk about the derived changes without including the homologous features of the sex differences themselves. Why not just grant that they are modified homologies, and include them as part of human nature?

Of course, human nature is just a term that we can define and use as we see fit. We can choose what we want the semantics of the term to be, and even use it differently in different contexts as long as we are explicit about what we mean. For example, we could restrict it to only universals (not particularly useful, in my view), or we could include human variation as part of human nature. What *does* matter, however we choose to use the term, is getting the biology right. You'd be mistaken if you thought, for example, that there is a set of modules that we can snap onto a chimp brain and make it human. This doesn't mean that chimp and human brains are the same, or that one is just a pumped-up version of the other. Nor does it mean, as some have suggested, that we should make an a priori decree of "no new modules" in explaining human-chimp differences (Donald, 2001). As we've seen, partitioning old from new in evolution isn't simple or easy, but that doesn't mean that there *isn't* change and therefore novelty. It's just that whatever new modules humans have are either modified versions of modules present in the chimp-human common ancestor (CHCA) or modified duplicates of ancestral modules (inparalogs). And we've also seen examples of how evolutionary changes can lead to new interactions between specialized mechanisms, as in the case of mindreading. In dynamical systems, altering interactions can be a game-changer.

Interactions between mechanisms and changes within systems of interacting parts make attribution of evolutionary trajectories to single causes problematic. As I mentioned, there are many searching for the holy grail of human-chimp differences—in some cases, postulating one new ability that took the chimp-human ancestral brain in a new direction, such as language, symbols, culture, or cooperation (Deacon, 1997; Donald, 2001; Penn et al., 2008; Tomasello, 1999). In my view, *all* of these things are likely to be part of the story, because all of these things do exist in humans. However, almost all of them probably interacted in their mutual evolution. As we saw in chapter 8, this makes it difficult to think in a chicken-and-egg way about which came first. This doesn't mean that some causes didn't come before others. But it also doesn't mean that first trait A changed and stopped, then B changed and stopped, then C changed and stopped; things coevolve. Indeed, there is plenty of evidence that evolution within our lineage has been a kind of accelerating, ratchet-like process with many features of human behavior, cognition, and social organization

changing in a sort of crescendo, with the volume still getting ever louder. Let's briefly consider the question of what this self-feeding evolutionary trajectory might have looked like.

---

Our earliest known ancestors, following the split of the human and chimpanzee lineages around seven million years ago, appear to have been little different than the likely CHCA, except for the fact that they had begun to rely on bipedal locomotion (Richmond & Jungers, 2008). In particular, their brains were similar in size to those of chimps. Hominin brain size didn't begin to increase drastically beyond the chimp range until nearly four million years later (Lee & Wolpoff, 2003; Striedter, 2005; Wood & Collard, 1999). Around the time that anthropologists mark as the origin of our genus, *Homo*, many changes began to occur. Our brains got bigger, our bodies got bigger, and our lifespans increased. Size dimorphism between males and females decreased, suggesting increased male investment in offspring and greater pair bonding. The earliest stone tools showing clear evidence of human manufacture begin to appear in the fossil record, and there is evidence they were used for hunting and/or scavenging. From this time period onward—around two million years ago—things began to slowly ratchet upward, picking up steam. Brains kept expanding, human artifacts became more complex, and populations expanded. During the last several hundred thousand years in Africa, human material culture and, presumably, cognitive sophistication accelerated in complexity, culminating in a migration of modern *Homo sapiens* out of Africa beginning approximately 60,000 years ago to populate the globe, largely replacing earlier human populations but in some cases mixing with them (Forster, 2004; McBrearty & Brooks, 2000; Reich et al., 2010).

What changes in brains and minds occurred during these last couple million years? Making inferences about this is clearly a difficult task but, contrary to the claims of extreme pessimists who claim we'll never know anything about it, we do have some sources of evidence to work with and will probably have more in the future than we do now. There is the stuff our ancestors left behind, including their artifacts, their modifications of the environment (shelters, fires, dead animal parts), and their bodies, mostly as fossils but sometimes with some DNA intact, which has been used in recent studies of so-called Neanderthal and Denisovan people, (partly) extinct lineages of ancestral hominins (Reich et al., 2010). And of course we have *our* DNA, which, being a product of the evolutionary process, bears at least some hints of what happened in those two million years and can be compared to the DNA of living chimpanzees, other primates, and the extinct Neanderthal and Denisovan lineages to make inferences about evolutionary changes (Khaitovich et al., 2006). We can also

study variation within modern human DNA to make inferences about migration patterns and the effects of recent selective events such as the appearance of agriculture, novel diseases, and other changing features of our environments (Novembre & Di Rienzo, 2009).

Genomic analyses are the newest part of this toolkit, and we are only beginning to understand the implications of genetic and epigenetic changes for evolution in brain design. However, what we can deduce from fossils and material artifacts suggests that multiple human traits were changing in a kind of evolutionary cascade. Anthropologist Hillard Kaplan and his colleagues have proposed such a self-feeding cascade scenario in which multiple features of human behavior, psychology, physiology, and interaction with the environment coevolved, each feeding back on the other (Kaplan et al., 2000). A key feature of their model is a move to complex foraging strategies that were not characteristic of earlier hominins or the CHCA: in particular, greater reliance on meat, foraged foods, and difficult-to-access foods like tubers. There is evidence for a shift toward more meat in the diet in the transition mentioned above. Meat is nutritionally dense, especially in materials like lipids that are important for brain growth. But hunting also requires sophisticated cognitive abilities in order to track, outsmart, and capture game. Thus, meat-eating could have coevolved with larger brain sizes in a mutually reinforcing cascade: Bigger brains mean better hunters; better hunters bring in more meat to build bigger brains. And an increased lifespan—in particular, a longer juvenile period—increases both the amount of time to grow a brain and the period of learning and plasticity that reaction norms can use to build brains based on experience and practice of hunting techniques.

Concurrently, sexual dimorphism data suggest that human males were investing more in their children—again, part of the spiral, because investing both food and training in offspring would help with brain growth and skill acquisition, leading to fitness gains for fathers. There is a social element as well, since hunting big game requires cooperation, therefore selecting for elements of social cognition such as theory of mind and a moral sense, as well as mechanisms enabling cultural transmission of hunting skills, communication skills (language), and norms of cooperation and coordinated activity. Big brains would, of course, help with this—and the additional neural real estate created by brain expansion could set the stage for the evolution and shaping of new mechanisms dedicated to things like language and skill acquisition. This was all, according to Kaplan and colleagues, part of a big virtuous circle.

What role does brain expansion play in this scenario? The easy answer is something like: "Brains got bigger, we got smarter, end of story." That is at some level a description of what happened, of course. But many people think that what this means is that you take the "general-purpose computational device" of the neocortex and just

add more. If you take general-purpose intelligence and add *more* general-purpose intelligence, then we'll get smarter at everything, right?

This is not, however, the only scenario. Neuroscientist Jon Kaas has pointed out that as brains get bigger, architectural considerations force them to become more modular, in at least one sense of the word (Kaas, 2000). In network theory, modularity refers to degree to which regions in the network have relatively more local within-region connections than distant between-region connections. A less modular network is one in which every part of the network is equally well connected, and a more modular one has dense "neighborhoods" connected by a few sparse highways. Imagine starting with a small sphere of neurons in which every neuron is connected to every other neuron. Now imagine expanding the sphere; as the number of neurons increases (as the cube of the width of the sphere, since it is a volume), it will become exponentially more difficult to keep every neuron attached to every other. As Kaas points out, the sheer volume of axons—neural wires—required to keep everything connected rapidly becomes intractable as brain size increases, forcing brain areas to become increasingly modular, or parcellated, with respect to each other. Consistent with this, evidence suggests that human brains have a relatively high degree of modularity, in this network sense (Bullmore & Sporns, 2009), and comparative work shows that species with larger brains tend to have greater differentiation of the expanded brain areas (e.g., cortex; Kaas, 1989; Striedter, 2005).

Does this mean that when you expand brains you instantly get new, functionally differentiated brain modules? No. However, it suggests that the stage is set for functional differentiation to occur—through reaction norms operating in duplicated, paralogous brain tissue, and eventually, through natural selection tuning those reaction norms. Indeed, one must be careful about the chicken and the egg here. One scenario is that mutations expand brain size and *then* modularization occurs. But if Kaas's proposal is correct, then mutations expanding brain size could be selected for *because* of the increased modularization they enable. Many brain scientists believe that modules are shaped by increasing expertise in specialized skills, like reading words, seeing faces, or playing tennis (Dehaene & Cohen, 2007; Karmiloff-Smith, 1992; Johnson & Vecera, 1996; Gauthier & Nelson, 2001). Thus, making new modular space available could have been selected for because of the new skills it enabled in hunting, cooperation, communication, social learning, and more.

Consider too that many of the changes in human minds over the last two million years would have involved tweaks and additions to existing brain architecture. It's not necessarily the case that older mechanisms are frozen in time, and newer mechanisms just slapped onto them—contrary to the popular view of an ancient, unmodified reptile brain with newer parts added later (MacLean, 1990; see Striedter, 2005, for a critique). We know, for example, that areas involved in speech perception

and production have homologs in other primates that do not speak, and yet there is evidence that these areas have been modified in humans, plausibly to enable spoken language (Buxhoeveden et al., 2001; Friederici, 2009; Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Rilling et al., 2008). Part of this can be due to new or altered interactions with other systems—for example, prefrontal white matter in humans, which plays a role in communication between different cortical areas, is expanded compared to other primates, suggesting selection for greater interactivity between cortical sub-processes (Schoenemann et al., 2005).

As I've stressed, interactions between systems, including and entailing modifications to preexisting systems, are likely to be crucial in explaining the evolution of modern human cognition. To take just one example, mindreading is likely to play a role in phenomena ranging from social learning, to communication, to cooperation. This means that selection for each of these things would, or could, *also* entail selection for more sophisticated mindreading skills. And selection of mindreading because of one fitness-beneficial function—say, the ability to infer goals in the context of learning how to make and use a tool—could have byproduct benefits, such as improving language acquisition or changing the shape of the adaptive landscape for some other skill, making it easier to evolve. Advances in neuroscience are making us increasingly aware that even though we might find brain regions that are differentially activated in a task—the bread and butter of recent brain mapping—we can also see that in everything we do, multiple regions are active and operating in a coordinated fashion. Although this is sometimes taken as evidence against modularity, it seems to me to be evidence for the coordinated interaction of multiple, specialized systems in any given task—the hallmark of flexibility through division of labor.

What this means for us is that in the search for what makes humans cognitively unique, we must take seriously the possibility of multiple changes: some small, some large, some to older systems, and some to newer ones. Many have suggested that two million years isn't much time for evolutionary change to occur, and that the genetic changes in that time have been too modest to allow for many or perhaps even any new functional specializations (Donald, 2001; Ehrlich, 2000; Tomasello, 1999). But the lessons of evo-devo suggest that rearranging developmental building blocks to create genuinely new things with relatively modest genetic changes is a common mode of evolution, and one that we should take seriously in thinking about brain evolution.

Because we don't yet have agreement on the nature of the mechanisms underlying human cognition, it's dicey to attempt a catalog of things about human cognition that we'll need to explain. But, in the spirit of putting hypotheses on the table, here is a crude and incomplete catalog of some of the mechanisms, processes, and abilities that appear to be part of the human cognitive package that likely coevolved in an evolutionary cascade.

Compared to chimps, we appear to have superior mindreading abilities, both in terms of the types of mental state representations we can entertain (John believes in God; Bill is trying to fix the carburetor) and the contexts in which we can entertain them (cooperation, inferring meaning in spoken language, parsing goals in skill learning). We have superior abilities of causal cognition, both in terms of the causal inferences we can bring to a learning problem (if I push the stick this way in the tube, the peanut will come out, but if I push the stick that way, it will fall into this hole before it comes out the other side), and the kinds of causal models we are able to learn from experience (formal physics, baking cakes). We have spoken language, including the ability to rapidly acquire lexical items, narrow down their meanings through a "gavagai" process, quickly learn grammatical rules in childhood, and apply these both to communicate with others and to acquire information from them through language that otherwise would have been very difficult to learn. Languages involve symbol systems that allow us to form representations that we could not form before, such as counting systems, the equations of formal mathematics, and the symbol systems of formal logic and its informal equivalents in natural grammars, that can both be used to reach new conclusions and store and transmit information that could previously not be stored or transmitted, much less conceived of. We have a variety of reasoning mechanisms that interface with language, such as the ability to construct metaphors and use them to reach new conclusions via metaphoric mapping, and a host of other higher-level reasoning besides. We are able to cooperate with others on scales never before seen in mammals, in contexts ranging from democratic voting to armed conflict, requiring mechanisms that alter our behavior to overlook immediate rewards or costs for larger gains. It's possible that we possess evolutionarily new (though derived) emotions and motivations, including specialized social emotions such as embarrassment and compassion. We are able to imitate and learn from others in ways that other animals cannot, and we are able to accumulate complex skills and representations that ratchet upward in complexity through processes of cultural evolution, which requires mechanisms that enable such complexity to be preserved without degradation, including, in all likelihood, mechanisms of language, theory of mind, and the ability to form complex causal models and recipes. And these are just the mechanisms and processes that I've reviewed in this book; there are almost certainly others.

---

Explaining humans needn't and shouldn't be the only goal of evolutionary psychology. We are just one species among many, and from a biological perspective, understanding how our minds work is no more or less important than understanding the

mind of a buzzard or a bumblebee. It is natural, of course, that we intrigue ourselves, and the desire to understand human nature isn't going away any time soon—nor should it. But to arrive at a biologically correct model of human nature, we're going to have to drop some bad habits.

First, we're going to have to stop denying the existence of human nature just because it can be waved away with some semantic sleight of hand. Yes, humans are variable and can't be cleanly defined by a checklist of necessary and sufficient features. Nevertheless, our species is a thing, a big wobbly cloud that is different from the population clouds of squirrels and palm trees. To understand human minds and behaviors, we need to understand the properties of our own cloud, as messy as it might be.

Second, we're going to have to put aside cartoon models of human nature and confront the biological facts. Evolutionary psychologists, for example, have been vilified for emphasizing the importance of evolutionarily recent timeframes—the Pleistocene—in shaping our evolved design. The point is well taken that *most* of human nature is in fact, far older: things like sociality, maternal care, male-female differences, our visual systems, emotional systems, motor systems, and much more. Recent events have undoubtedly shaped our psychology in important ways and even created some genuinely new skills, but human nature didn't just burst forth on the Pleistocene savannah. It's time to put this debate to rest and realize that the study of human evolution is the study of *everything* that has shaped us, all the way back—including, but not limited to, what happened in the Pleistocene.

Finally, psychology needs to shift the frame from seeing humans as a sui generis phenomenon to seeing us as one tip of one branch on the evolutionary tree. We tend to study human psychological abilities in isolation, without considering homology and the ways in which our abilities have arisen from older ones via descent with modification. Debates about human sex differences, for example, frequently pay little or no attention to the fact that they are largely homologous with those seen in other primates. Although it's possible that they have been completely eliminated in humans, it would be a surprise.[2] Similarly, abilities such as general-purpose learning are sometimes invoked to explain uniquely human traits like language when those abilities are also present in other primates that don't have language. This doesn't mean that the abilities in question couldn't have been modified in us, but if so, this needs to be part of the hypothesis: *How* has the ability in question been modified to enable something in humans that is not seen in other species?

[2] Among other things, this implies that the hypothesis of evolved sex differences in humans, including at least some psychological sex differences, is favored by phylogenetic parsimony. The burden of proof lies with those who argue that sex differences have been entirely eliminated in us.

# CONCLUSION
# POSSIBILITIES

As an evolutionary psychologist, one is bombarded on a nearly daily basis with pessimism. Pummeled with it, really. Much of this pessimism has an a priori flavor as in: We already know. We already know how development occurs. We already know as much as we'll ever know about the evolutionary past. We already know that what you're proposing is impossible. And then there is the even more depressing brand of pessimism that says we'll *never* know. The brain is too complicated. The evolutionary past is too long ago. Society, development, and humans are just too complex to be explained by the kinds of explanations you're proposing.

Maybe. The possibility that some things are simply unknowable is, of course, one that every scientist should entertain. But if a question's unanswerability at the time it was asked were grounds for rejecting it, some of the most interesting discoveries in science wouldn't have happened. And as for the idea that we already know as much as we're ever going to know, does anybody seriously think that if we could somehow call a timeout and write down everything that is currently known about the mind, that we'd be done?

Many would agree with these points, but would nevertheless argue that there is something fundamentally wrong with attempting to understand the mind as composed of adaptations. The reasons stretch from a priori claims about the mere possibility of adaptations in the brain (in the cortex especially) to the claim that hypotheses about how selection has shaped the brain are essentially unknowable, because they are unfalsifiable.

I hope I've convinced you that the first reason—higher-level brain adaptations are impossible or highly unlikely—is a nonstarter. At minimum, despite claims to the contrary, we just don't know enough about brain development to say confidently that it's true. But more than that, there are strong reasons to think that some version of the thesis that the mind is composed of adaptations, all the way up, *is* true. I've reviewed many of them here, both theoretical and empirical: There are good reasons to believe that all cognition is carried out via designed interactions between

specialized systems. This doesn't mean, of course, that any current model that we have of how this occurs is sure to be correct. It just means that any reasonable theory of brain evolution needs to include it as a possibility.

The second argument for abandoning adaptationism is the "just-so story" position, which holds that answers to adaptationist questions are unknowable, so we should stop asking them. Or at best we should ask them with extreme caution, as if we're performing open heart surgery or handling nuclear materials. Not only do I (mostly) disagree, I'll go one step further and propose the following bit of heresy. When it comes to the mind and the matrix of causes and effects within which it's embedded—evolutionary, developmental, neural—the more hypotheses, the better. The problem is not that we have too many hypotheses; it's that we don't have enough. We don't yet know which ones we will and won't eventually be able to test, and we as yet have a poor idea of the scope of possible developmental and computational designs that the evolutionary process can create. So, bring them on.

Before you start sharpening your Occam's razor, let me be clear. Some hypotheses about the mind are indeed quite implausible and should be accorded a low prior probability (though I suspect that we overestimate our abilities to assign priors of this kind). Some are even absurd; I've read many in peer-reviewed journals that caused me to laugh out loud.[1] But so what? Shall we have a hypothesis police that prevents people from even thinking these things? In medicine, economics, and many sciences, hypotheses that strain credulity are proposed. You could argue that these can cause genuine harm, and you might point to cases like the vaccination theory of autism or the invisible hand theory of markets. Some would (and do) argue that evolutionary hypotheses have a special gravitas that makes them the scientific equivalent of an automatic weapon. While I have my doubts about this as an empirical matter—I really don't know of any evolutionary hypotheses that have done as much damage as economic or medical ones—I don't disagree that it's important to consider the social consequences of scientific theorizing. That's not what I'm arguing.

What I am arguing, first, is what Bayesians argue about model-building: If you don't have a hypothesis within your set of possibilities, then no evidence can speak for or against it. To be sure, hypotheses should be formulated wisely, on the basis of plausible assumptions. But what I've tried to argue in this book is that the scope of what's plausible in evolution and development is much wider than the currently rather impoverished set of models in psychology. These include models that we already know must be too simple: the model that mental mechanisms come in two

[1] Define a just-so story as a story that *could* account for the data if it were true—a big if. Are there really more just-so stories in evolutionary psychology than in social science in general?

flavors (specialized and general purpose), the model of development as innate knowledge plus general-purpose learning, or the model of cognition as emerging purely from statistical learning over perceptual inputs. The taboo on adaptationist thinking imposed by Gould and Lewontin has had a chilling effect on evolutionary theorizing, one that is now plausibly hurting more than helping. When their critique was first introduced over 30 years ago, their basic points—that not everything is an adaptation and that evolutionary hypotheses are hard to test—were well taken. But at this point, they've been sufficiently absorbed into the scientific community that ordinary mechanisms of adjudicating hypotheses—experiments, peer review, and the like—should be allowed to do their work.

I've tried to present a case that many of the hypotheses currently on the table to explain human cognition are likely to be orders of magnitude too simple. It's true, for example, that language has to be an important part of explaining human uniqueness—but it's not enough. The same goes for cultural transmission, the ability to manipulate symbols, theory of mind, and prosociality. Each of these is likely to be an important piece of the story, but not the only piece. And even then, the pieces have sub-pieces. Something like what we can call with three little words, theory of mind, is in fact likely to be composed of many smaller pieces, each with its own evolutionary raison d'être. This is the second sense in which I mean that we are going to need more hypotheses, not fewer.

Parsimony has many definitions, but a reasonable gloss for the way it is generally used in science is something like this: When there are multiple hypotheses consistent with observed data, prefer the simplest one. This follows from William of Ockham's injunction not to multiply explanatory entities beyond necessity. The intention, presumably, is to rein in theoretical flights of fancy and keep our theories as close as possible to the facts. But in order to follow this rule, you need to know what "necessity" is. In the case of mental evolution, it dictates that we should postulate no more causes or mechanisms than we need to explain the observed facts of cognition and behavior. This all seems well and good until we realize, as is the case for much of human cognition, that we often have little to no idea what the true underlying amount of complexity necessary to support our hypothesis really is. Though we might think otherwise, we are largely in the dark about what developmental designs are easier or harder for evolution to produce, or what truly makes one process or mechanism simpler than another. At best, we rely on formal models, which are almost always recognized to be simpler by some difficult-to-estimate amount than the actual stuff they are modeling. Much worse and much more commonly, we rely on intuitions, such as that a hypothesis is simpler if it requires fewer words or fewer concepts to describe it. Under such an assumption, simple explanations like learning or innateness are nearly guaranteed to win, sweeping under the rug questions of how many

lines of biological code—genes, neurons, bits of information, etc.—you'd need to learn a thing or specify it innately.

As many evolutionary biologists have pointed out, even though evolution may produce seemingly ingenious solutions to adaptive problems, they are rarely produced in the simplest possible way. This might seem like bad news for us epistemologically, and I think that in a sense it is. But wishing for a one-to-one mapping between how easily a concept can be stated and how likely it is to actually explain a phenomenon won't necessarily make it so. Instead, we might do well to abandon naïve hopes about parsimony as an all-purpose truth-finding tool and realize that what matters is arriving at the correct explanation of the mind, not the simplest one. Comparing complexity alone will not be enough to get us to the correct explanation; we must compare hypotheses, ultimately, based on what they can and can't explain, no matter how complex they are. All of this, I hope, serves to underline the perhaps depressing point that the more biologically realistic our theories become, the less complete our current explanations of the mind appear to be.

What I've tried to argue in this book, then, is that we should look at the mind with a bigger lens, one that isn't narrowed to allow for only a few possibilities, but widened in its scope and sharpened in its focus to capture what's really there, not what we'd like to see. There are in fact many possibilities for mind design, at many different levels of organization, not just the few suggested by major schools of thought in contemporary psychology and philosophy of mind. And the conceptual tools that we currently use are mostly too blunt and too old, inherited from distant, even pre-biological conceptions of the mind. When these tools fail to fit biological, psychological, and cultural reality, we should discard them and replace them with new concepts designed to capture the phenomenon in question and no other. Although the theoretical framework I've proposed in this book is by no means exhaustive, here is a brief summary of the proposals I've made along these lines.

I've suggested that we reconsider the igloo model of the mind as composed of a layer of crunchy domain-specific modules surrounding a soft domain-general center. Instead, we should be prepared for a mind that contains a diversity of different parts and processes, evolved through descent with modification in an evolutionary mosaic. In our attempt to develop and test theories of these parts and processes, we should be mindful of the first law of adaptationism: It depends. Everything in biology depends on history and circumstance, so the human urge to compartmentalize everything into neat, logical categories needs to take a backseat to the facts of life.

As a means of formalizing proposals about mental evolution, I've suggested that we think in terms of possibility spaces: spaces of possible genotypes, spaces of possible phenotypes, and the probabilistic functions that map between them. I've proposed that we think about those mapping functions using the biological concept of

a reaction norm, and that we broaden that concept to include not just classical or one-to-one reaction norms, but also open reaction norms that have been selected because of their outcomes in massively multidimensional spaces. Importantly, and especially in the case of humans, these include reaction norms designed to handle cultural inputs including language, social norms, and artifacts.

The model of cognition that I have proposed blends elements of traditional modular, symbol-representation views with elements of parallel distributed processing and dynamical systems views. In particular, there is good evidence that the whole of cognition is achieved through a collaborative computational division of labor. Cognition is synergistic, contextually driven, and dynamical—but by design. That design is engineered into the reaction norms that build brains through emergent processes of self-organization. Key to this view is the idea that adaptations instantiate inductive bets: Adaptations are designed to interface with the world and with each other by the successes and failures of the past, but those design bets only pay off when their felicity conditions hold true. Inductive bets are therefore probabilistic and exist at all levels of the architecture, from the bets instantiated in reaction norms and learning rules to the bets instantiated in developed phenotypes.

I've suggested, as have many psychologists before, that the mind's mechanisms—and therefore evolution—are responsible for the phenomenon that we experience as meaning. Meaning doesn't exist objectively in the world, but is rather painted onto it by the mind. Our minds semantically colorize the world around us in many ways, from the meanings painted onto the world by our perceptual systems to cultural meanings, such as the meanings of words, practices, and concepts, that our culture acquisition mechanisms enable us to acquire. And in semantic colorization, the phenomenon of *enabling* is key: Without evolved mechanisms, no culturally transmitted meaning would be possible at all; nor would the crucially important phenomenon of ratcheting cultural evolution, which allows complexity and design to accumulate in the cultural entities that our minds create, produce, transmit, and acquire.

Finally, I have proposed that we stop thinking about human responses to evolutionary novelty as the purely coincidental interaction of domain-general mechanisms with a haphazardly unprecedented world, and start thinking about them as products of coevolutionary interactions that are in many ways—including most of the ways that enable adaptive responses to novelty—non-coincidental. They are non-coincidental on both sides of the mind-world equation. Our minds respond adaptively to novelty if and only if the novelty is not truly novel, at least in some aspects: It must satisfy the felicity conditions of some inductive bets that build and operate our minds. And in cases of responses to novelty that are cultural in origin, the mind-world fit comes from both sides of the dynamical system, because minds and their cultural products coevolve. Cultural phenomena such as norms and artifacts often evolve because they

satisfy the felicity conditions of mental adaptations, resulting in products designed to fit the mind. Thus, neither pure general-purposeness nor future-seeing teleology need be part of this equation. Mechanisms can have functions, including functions with varying types and degrees of specialization, and culture can evolve. No contradiction is entailed.

In my view, this set of concepts, or something like it, has the potential to bridge the gap that I described at the beginning of the book: the gap between the realization that the mind is the product of evolution and the total lack of agreement about *what* in the mind is the product of evolution. The fashion, at the moment, is to try to partition the mind into those things that are the products of evolution, and those that aren't. What I've tried to argue is that on a truly interactionist view, there is nothing that isn't: Everything in the mind is the product of interactions between evolved, designed, developmental systems and the world in which they develop. On such a view there will ultimately be no gap—nothing extra left to explain that doesn't have evolutionary origins.

I think we already know enough to expect this will be the case. The hard part, of course, is everything that lies in between: figuring out the stuff that intervenes between evolutionary history and phenotypic outcomes. To do this, there is no substitute for looking at the thing itself and asking what you want to know about it, and then designing your inquiries to answer those questions and no others. In our case, we're interested in knowing how one strange and complex process—evolution—shaped another strange and complex process—development—which shaped another strange and complex process—thought. We should be prepared for the possibility that the explanation for it all will end up being strange and complex as well.

# REFERENCES


Adamson, R. E. (1952). Functional fixedness as related to problem solving: A repetition of three experiments. *Journal of Experimental Psychology, 44*(4), 288–291.

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 21–62.

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature, 393*(6684), 470–474.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature, 372*(6507), 669–672.

Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting: Choice over time. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 57–92). New York: Russell Sage Foundation.

Alberch, P. (1982). Developmental constraints in evolutionary processes. In J. T. Bonner (Ed.), *Evolution and development: Report of the Dahlem workshop on evolution and development* (pp. 313–332). New York: Springer-Verlag.

Allman, J. M. (2000). *Evolving brains*. New York: Scientific American Library.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277.

Anderson, J. R. (1996). *The architecture of cognition*. New York: Psychology Press.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*(4), 245–266.

Andersson, M., & Iwasa, Y. (1996). Sexual selection. *Trends in Ecology and Evolution, 11*(2), 53–58.

Ariew, A. (1999). Innateness is canalization: In defense of a developmental account of innateness. In V. G. Hardcastle (Ed.), *Where biology meets psychology: Philosophical essays* (pp. 117–138). Cambridge, MA: MIT Press.

Aunger, R. (2002). *The electric meme: A new theory of how we think*. New York: Free Press.

Austin, J. L. (1962). *How to do things with words.* New York: Oxford University Press.

Ausubel, F. M. (2005). Are innate immune signaling pathways in plants and animals conserved? *Nature Immunology, 6*(10), 973–979.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. New York: Oxford University Press.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104*(1), 54–75.

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants. *Cognition, 23*(1), 21–41.

Baillargeon, R., & Hanko-Summers, S. (1990). Is the top object adequately supported by the bottom object? Young infants' understanding of support relations. *Cognitive Development, 5*(1), 29–53.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition, 20*(3), 191–208.

Baird, J. A., & Baldwin, D. A. (2001). Making sense of human behavior: Action parsing and intentional inference. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 193–206). Cambridge, MA: MIT Press.

Balda, R. P., & Kamil, A. C. (1992). Long-term spatial memory in Clark's nutcracker, *Nucifraga columbiana*. *Animal Behaviour, 44*(4), 761–769.

Baldwin, C. Y., & Clark, K. B. (2000). *Design Rules: Vol. 1. The power of modularity.* Cambridge, MA: MIT Press.

Baldwin, D. A., & Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences, 5*(4), 171–178.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*(3), 708–717.

Bandura, A. (1977). *Social learning theory.* New York: General Learning Press.

Bardy, S. L., Ng, S. Y. M., & Jarrell, K. F. (2003). Prokaryotic motility structures. *Microbiology, 149*(2), 295–304.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *Journal of Physiology, 119*(1), 69–88.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46.

Baron-Cohen, S., Tager-Flusberg, H., & Lombardo, M. V. (Eds.). (2013). *Understanding other minds: Perspectives from developmental social neuroscience.* New York: Oxford University Press.

Barrett, H. C. (2001). On the functional orgins of essentialism. *Mind & Society, 2*(1), 1–30.

Barrett, H. C. (2005a). Adaptations to predators and prey. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 200–223). New York: Wiley.

Barrett, H. C. (2005b). Enzymatic computation and cognitive modularity. *Mind & Language, 20*(3), 259–287.

Barrett, H. C. (2006). Modularity and design reincarnation. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Vol. 2. Culture and cognition* (pp. 199–217). New York: Oxford University Press.

Barrett, H. C. (2007). Development as the target of evolution: A computational approach to developmental systems. In S. W. Gangestad & J. A. Simpson (Eds.), *The evolution of mind: Fundamental questions and controversies* (pp. 186–192). New York: Guilford.

Barrett, H. C. (2009). Where there is an adaptation, there is a domain: The form-function fit in information processing. In S. M. Platek & T. K. Shackelford (Eds.), *Foundations in evolutionary cognitive neuroscience* (pp. 97–116). Cambridge, UK: Cambridge University Press.

Barrett, H. C. (2012). A hierarchical model of the evolution of human brain specializations. *Proceedings of the National Academy of Sciences, 109*(Suppl. 1), 10733–10740.

Barrett, H. C., & Behne, T. (2005). Children's understanding of death as the cessation of agency: A test using sleep versus death. *Cognition*, *96*(2), 93–108.

Barrett, H. C., & Broesch, J. (2012). Prepared social learning about dangerous animals in children. *Evolution and Human Behavior, 33*(5), 499–508.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113*(3), 628–647.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., & Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences. 280*(1755), 20122654.

Barrett, H. C., Laurence, S., & Margolis, E. (2008). Artifacts and original intent: A cross-cultural perspective on the design stance. *Journal of Cognition and Culture, 8*(1–2), 1–22.

Barrett, H. C., Stich, S., & Laurence, S. (2012). Should the study of *Homo sapiens* be part of cognitive science? *Topics in Cognitive Science, 4*(3), 379–386.

Bartlett, J. C., & Searcy, J. (1993). Inversion and configuration of faces. *Cognitive Psychology, 25*(3), 281–316.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences, 108*(18), 7641–7646.

Bateson, G. (2000). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology.* Chicago: University of Chicago Press.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(1), 59–78.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295–307.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41*(2), 328–337.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312.

Belsky, J., Steinberg, L. D., Houts, R. M., Friedman, S. L., DeHart, G., Cauffman, E., & The NICHD Early Child Care Research Network. (2007). Family rearing antecedents of pubertal timing. *Child Development, 78*(4), 1302–1321.

Ben-Shahar, Y., Robichon, A., Sokolowski, M. B., & Robinson, G. E. (2002). Influence of gene action across different time scales on behavior. *Science, 296*(5568), 741–744.

Bernasco, W. (2009). Foraging strategies of Homo criminalis: Lessons from behavioral ecology. *Crime Patterns and Analysis, 2*(1), 5–16.

Bilder, R. M., Sabb, F. W., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., & Poldrack, R. A. (2009). Cognitive ontologies for neuropsychiatric phenomics research. *Cognitive Neuropsychiatry, 14*(4–5), 419–450.

Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition, 107*(3), 1018–1034.

Bjorklund, D. F., Ellis, B. J., & Rosenberg, J. S. (2007). Evolved probabilistic cognitive mechanisms: An evolutionary approach to gene x environment x development interactions. In R. V. Kail (Ed.), *Advances in Child Development and Behavior* (Vol. 35, pp. 1–39). San Diego, CA: Academic Press.

Blackmore, S. (1999). *The meme machine.* New York: Oxford University Press.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58*, 47–73.

Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K., & Gilad, Y. (2008). Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genetics, 4*(11), e1000271.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247.

Block, N. J., Flanangan, O., & Güzeldere, G. (Eds.). (1997). *The nature of consciousness: Philosophical debates*. Cambridge, MA: MIT Press.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), B25–B31.

Blumstein, D. T., & Daniel, J. C. (2005). The loss of anti-predator behaviour following isolation on islands. *Proceedings of the Royal Society B: Biological Sciences, 272*(1573), 1663–1668.

Boesch, C. (1994). Cooperative hunting in wild chimpanzees. *Animal Behaviour, 48*(3), 653–667.

Bogin, B. (1999). *Patterns of human growth* (2nd ed.). New York: Cambridge University Press.

Bolhuis, J. J., Brown, G. R., Richardson, R. C., & Laland, K. N. (2011). Darwin in mind: New opportunities for evolutionary psychology. *PLoS Biology, 9*(7), e1001109.

Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognitive Psychology, 44*(4), 388–426.

Boroditsky, L. (2011). How language shapes thought. *Scientific American, 304*(2), 62–65.

Boto, L. (2010). Horizontal gene transfer in evolution: Facts and challenges. *Proceedings of the Royal Society B: Biological Sciences, 277*(1683), 819–827.

Bowlby, J. (1969). *Attachment and Loss: Vol. 1. Attachment.* New York: Basic Books.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Boyd, R., & Richerson, P. J. (1988). An evolutionary model of social learning: The effects of spatial and temporal variation. In T. R. Zentall & B. G. Galef Jr. (Eds.), *Social learning: Psychological and biological perspectives* (pp. 29–48). Hillsdale, NJ: Lawrence Erlbaum.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*(3), 171–195.

Boyd, R., & Richerson, P. J. (1996). Why culture is common but cultural evolution is rare. In W. G. Runciman, J. Maynard Smith, & R. I. M. Dunbar (Eds.), *Proceedings of the British Academy: Vol. 88. Evolution of social behaviour patterns in primates and man* (pp. 73–93). Oxford, UK: Oxford University Press.

Boyer, P., & Barrett, H. C. (2005). Domain specificity and intuitive ontology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 96–118). New York: Wiley.

Braatsch, S., & Klug, G. (2004). Blue light perception in bacteria. *Photosynthesis Research, 79*(1), 45–57.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Brakefield, P. M. (2006). Evo-devo and constraints on selection. *Trends in Ecology and Evolution, 21*(7), 362–368.

Braun, J. (2003). Natural scenes upset the visual applecart. *Trends in Cognitive Sciences, 7*(1), 7–9.

Bridgeman, B., Kirch, M., & Sperling, A. (1981). Segregation of cognitive and motor aspects of visual function using induced motion. *Perception & Psychophysics, 29*(4), 336–342.

Breuker, C. J., Debat, V., & Klingenberg, C. P. (2006). Functional evo-devo. *Trends in Ecology and Evolution, 21*(9), 488–492.

Broesch, J., Barrett, H. C., & Henrich, J. (2014). Adaptive content biases in learning about animals across the life course. *Human Nature, 25*, 1-19.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193–217.

Bryant, G. A., & Barrett, H. C. (2007). Recognizing intentions in infant-directed speech: Evidence for universals. *Psychological Science, 18*(8), 746–751.

Buller, D. J. (2005). *Adapting minds: Evolutionary psychology and the persistent quest for human nature*. Cambridge, MA: MIT Press.

Buller, D. J., & Hardcastle, V. G. (2000). Evolutionary psychology, meet developmental neurobiology: Against promiscuous modularity. *Brain and Mind, 1*(3), 307–325.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience, 10*(3), 186–198.

Buxhoeveden, D. P., Switala, A. E., Litaker, M., Roy, E., & Casanova, M. F. (2001). Lateralization of minicolumns in human planum temporale is absent in nonhuman primate cortex. *Brain, Behavior and Evolution, 57*(6), 349–358.

Byers, J. A. (1997). *American pronghorn: Social adaptations and the ghosts of predators past.* Chicago: University of Chicago Press.

Byrne, R., & Whiten, A. (Eds.). (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans.* New York: Oxford University Press.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12*(1), 1–47.

Caceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., & Barlow, C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences, 100*(22), 13030–13035.

Cai, L., Dalal, C. K., & Elowitz, M. B. (2008). Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature, 455*(7212), 485–491.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Carey, S. (2009). *The origin of concepts.* New York: Oxford University Press.

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: Beyond object files and object tracking. *Cognition, 80*(1–2), 179–213.

Carnap, R. (1937). *The logical syntax of language* (A. Smeaton, Trans.). London: Routledge & Kegan Paul.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21*(2), 315–330.

Carpenter, M., Call., J., & Tomasello, M. (2005). Twelve- and 18-month-olds copy actions in terms of goals. *Developmental Science, 8*(1), F13–F20.

Carroll, S. B. (2005). *Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom*. New York: Norton.

Carroll, S. B. (2006). *The making of the fittest: DNA and the ultimate forensic record of evolution*. New York: Norton.

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell, 134*(1), 25–36.

Carroll, S. B., Grenier, J. K., & Weatherbee, S. D. (2005). *From DNA to diversity: Molecular genetics and the evolution of animal design* (2nd ed.). Oxford, UK: Blackwell.

Carruthers, P. (2006). *The architecture of the mind*. New York: Oxford University Press.

Cashdan, E. (1994). A sensitive period for learning about food. *Human Nature, 5*(3), 279–291.

Casler, K., & Kelemen, D. (2005). Young children's rapid learning about artifacts. *Developmental Science, 8*(6), 472–480.

Casler, K., & Kelemen, D. (2007). Reasoning about artifacts at 24 months: The developing teleo-functional stance. *Cognition, 103*(1), 120–130.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton, NJ: Princeton University Press.

Chalmers, D. J. (2010). *The character of consciousness*. New York: Oxford University Press.

Chang, D. H. F., & Troje, N. F. (2009). Characterizing global and local mechanisms in biological motion perception. *Journal of Vision, 9*(5), 1–10.

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B: Biological Sciences, 272*(1560), 267–275.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9*(2), 129–136.

Chater, N., & Oaksford, M. (2008). The probabilistic mind: Prospects for a Bayesian cognitive science. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for a Bayesian cognitive science* (pp. 3–32). New York: Oxford University Press.

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences, 106*(4), 1015–1020.

Chen, C., Burton, M., Greenberger, E., & Dmitrieva, J. (1999). Population migration and the variation of dopamine D4 receptor (DRD4) allele frequencies around the globe. *Evolution and Human Behavior, 20*(5), 309–324.

Cheney, D. L., & Seyfarth, R. M. (2007). *Baboon metaphysics: The evolution of a social mind*. Chicago: University of Chicago Press.

Chiappe, D., & MacDonald, K. (2005). The evolution of domain-general mechanisms in intelligence and learning. *Journal of General Psychology, 132*(1), 5–40.

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

Chomsky, N. (1972). *Language and mind*. New York: Harcourt Brace Jovanovich.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*(5), 489–509.

Chudek, M., Heller, S., Birch, S., & Henrich, J. (2012). Prestige-biased cultural learning: Bystander's differential attention to potential models influences children's learning. *Evolution and Human Behavior, 33*(1), 46–56.

Churchland, P. S., & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives, 4*, 343–382.

Claidière, N., & Sperber, D. (2007). The role of attraction in cultural evolution. *Journal of Cognition and Culture, 7*(1–2), 89–111.

Claidière, N., & Sperber, D. (2010). Imitation explains the propagation, not the stability of animal culture. *Proceedings of the Royal Society B: Biological Sciences, 277*(1681), 651–659.

Clark, H. H. (1996). *Using language.* New York: Cambridge University Press.

Clark, J. A. (2010). Relations of homology between higher cognitive emotions and basic emotions. *Biology & Philosophy, 25*(1), 75–94.

Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids: The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 507–522.

Coen, E. (1999). *The art of genes: How organisms make themselves.* New York: Oxford University Press.

Conover, D. O., & Schultz, E. T. (1995). Phenotypic similarity and the evolutionary significance of countergradient variation. *Trends in Ecology and Evolution, 10*(6), 248–252.

Cook, S. A. (1982). An overview of computational complexity. *Communications of the ACM, 26*(6), 400–409.

Cook, S., & Mineka, S. (1989). Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. *Journal of Abnormal Psychology, 98*(4), 448–459.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition, 50*(1–3), 41–77.

Coss, R. G., & Goldthwaite, R. O. (1995). The persistence of old designs for perception. In N. S. Thompson (Ed.), *Perspectives in Ethology: Vol. 11. Behavioral design* (pp. 83–148). New York: Plenum Press.

Crick, F. (1968). The origin of the genetic code. *Journal of Molecular Biology, 38*(3), 367–379.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*(2), 119–126.

Csibra, G. (2007). Action mirroring and action interpretation: An alternative account. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Attention and Performance XXII: Sensorimotor foundations of higher cognition* (pp. 435–459). New York: Oxford University Press.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition, 107*(2), 705–717.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148–153.

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), 1149–1157.

Culham, J. C., & Valyear, K. F. (2006). Human parietal cortex in action. *Current Opinion in Neurobiology, 16*(2), 205–212.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.

Dally, J. M., Emery, N. J., & Clayton, N. S. (2005). Cache protection strategies by western scrub-jays, *Aphelocoma californica:* Implications for social cognition. *Animal Behaviour, 70*(6), 1251–1263.

Dally, J. M., Emery, N. J., & Clayton, N. S. (2006). Food-caching Western scrub-jays keep track of who was watching when. *Science, 312*(5780), 1662–1665.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness.* New York: Harcourt.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23*(8), 3423–3431.

Dawkins, R. (1976). *The selfish gene*. New York: Oxford University Press.

Dawkins, R. (1997). *Climbing Mount Improbable*. New York: Norton.

de Gelder, B., & Rouw, R. (2000). Configural face processes in acquired and developmental prosopagnosia: Evidence for two separate face systems? *Neuroreport, 11*(14), 3145–3150.

de Haan, M., & Nelson, C. A. (1999). Brain activity differentiates face and object processing in 6-month-old infants. *Developmental Psychology, 35*(4), 1113–1121.

De Valois, R. L., & De Valois, K. K. (1988). *Spatial vision*. New York: Oxford University Press.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. New York: Norton.

Deaner, R. O., & Nunn, C. L. (1999). How quickly do brains catch up with bodies? A comparative method for detecting evolutionary lag. *Proceedings of the Royal Society B: Biological Sciences, 266*(1420), 687–694.

Decety, J., & Grèzes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences, 3*(5), 172–178.

Dehaene, S. (2009). *Reading in the brain: The science and evolution of a human invention*. New York: Viking.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron, 56*(2), 384–398.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*(1–2), 1–37.

DeLoache, J. S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development, 62*(4), 736–752.

Dennett, D. (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), *Minds, machines and evolution* (pp. 129–151). New York: Cambridge University Press.

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

Denton, D. (1982). *The hunger for salt: An anthropological, physiological and medical analysis*. New York: Springer-Verlag.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*(2), 281–299.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107–117.

Dickins, T. E., & Rahman, Q. (2012). The extended evolutionary synthesis and the role of soft inheritance in evolution. *Proceedings of the Royal Society B: Biological Sciences, 279*(1740), 2913–2921.

Diesendruck, G., Markson, L., & Bloom, P. (2003). Children's reliance on creator's intent in extending names for artifacts. *Psychological Science, 14*(2), 164–168.

Dindo, M., Thierry, B., & Whiten, A. (2008). Social diffusion of novel foraging methods in brown capuchin monkeys (*Cebus apella*). *Proceedings of the Royal Society B: Biological Sciences, 275*(1631), 187–193.

DiYanni, C., & Kelemen, D. (2008). Using a bad tool with good intention: Young children's imitation of adults' questionable choices. *Journal of Experimental Child Psychology, 101*(4), 241–261.

Donald, M. (2001). *A mind so rare: The evolution of human consciousness*. New York: Norton.

Duchaine, B. C., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology, 23*(5), 714–747.

Dugatkin, L. A. (1992). Sexual selection and imitation: Females copy the mate choice of others. *The American Naturalist, 139*(6), 1384–1389.

Dulai, K. S., von Dornum, M., Mollon, J. D., & Hunt, D. M. (1999). The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. *Genome Research, 9*(7), 629–638.

Duncker, K. (1945). On problem solving (L. S. Lees, Trans.). *Psychological Monographs, 58*(5, Whole No. 270).

Dupré, J. (2001). *Human nature and the limits of science.* New York: Oxford University Press.

Duranti, A. (1988). Intentions, language, and social action in a Samoan context. *Journal of Pragmatics, 12*(1), 13–33.

Ehrlich, P. R. (2000). *Human natures: Genes, cultures, and the human prospect.* Washington, DC: Island.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften, 58*(10), 465–523.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Emery, N. J. (2006). Cognitive ornithology: The evolution of avian intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1465), 23–43.

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature, 414*(6862), 443–446.

Emery, N. J., & Clayton, N. S. (2008). How to build a scrub-jay that reads minds. In S. Itakura & K. Fujita (Eds.), *Origins of the social mind: Evolutionary and developmental views* (pp. 65–97). Japan: Springer.

Emlen, S. T. (1975). The stellar-orientation system of a migratory bird. *Scientific American, 233*(2), 102–111.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., & Pääbo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. *Science, 296*(5566), 340–343.

Endler, J. A. (1986). *Natural selection in the wild.* Princeton, NJ: Princeton University Press.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.

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), 429–448.

Everett, D. L. (2012). *Language: The cultural tool.* New York: Random House.

Falconer, D. S. (1990). Selection in different environments: Effects on environmental sensitivity (reaction norm) and on mean performance. *Genetics Research, 56*(1), 57–70.

Farah, M. J. (1990). *Visual agnosia: Disorders of object recognition and what they tell us about normal vision.* Cambridge, MA: MIT Press.

Feldman, M. W., Aoki, K., & Kumm, J. (1996). Individual versus social learning: Evolutionary analysis in a fluctuating environment. *Anthropological Science, 104*(3), 209–232.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 10*(3), 279–293.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*(3), 477–501.

Fernald, R. D. (2000). Evolution of eyes. *Current Opinion in Neurobiology, 10*(4), 444–450.

Fessler, D. M. T. (2002). Reproductive immunosuppression and diet: An evolutionary perspective on pregnancy sickness and meat consumption. *Current Anthropology, 43*(1), 19–39.

Fessler, D. M. T. (2006). Steps toward an evolutionary psychology of a culture-dependent species. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Vol. 2. Culture and cognition* (pp. 61–77). New York: Oxford University Press.

Fessler, D. M. T., & Navarrete, C. D. (2003). Meat is good to taboo: Dietary proscriptions as a product of the interaction of psychological mechanisms and social processes. *Journal of Cognition and Culture, 3*(1), 1–40.

Fisher, R. A. (1919). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52*(2), 399–433.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). London: Sage.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Biology, 19*(2), 99–113.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology, 15*(5), 447–452.

Fodor, J. (1998). The trouble with psychological Darwinism. *London Review of Books, 20*(2), 11–13.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of computational psychology.* Cambridge, MA: MIT Press.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In J. L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (pp. 26–36). Cambridge, MA: MIT Press.

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior, 4*(5), 414–420.

Forster, P. (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359*(1442), 255-264.

Foster, K. W., & Smyth, R. D. (1980). Light antennas in phototactic algae. *Microbiological Reviews, 44*(4), 572–630.

Fragaszy, D., & Visalberghi, E. (2004). Socially biased learning in monkeys. *Learning & Behavior, 32*(1), 24–35.

Frankenhuis, W. E., & Barrett, H. C. (2013). Design for learning: The case of chasing. In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 171–196). Cambridge, MA: MIT Press.

Frankenhuis, W. E., & Panchanathan, K. (2011). Balancing sampling and specialization: An adaptationist model of incremental development. *Proceedings of the Royal Society B: Biological Sciences, 278*(1724), 3558–3565.

Frankenhuis, W. E., House, B., Barrett, H. C., & Johnson, S. P. (2013a). Infants' perception of chasing. *Cognition, 126*(2), 224–233.

Frankenhuis, W. E., Panchanathan, K., & Barrett, H. C. (2013b). Bridging developmental systems theory and evolutionary psychology using dynamic optimization. *Developmental Science, 16*(4), 584–598.

Franks, N. R., & Richardson, T. (2006). Teaching in tandem-running ants. *Nature, 439*(7073), 153.

Franz, M., & Matthews, L. J. (2010). Social enhancement can create adaptive, arbitrary and maladaptive cultural traditions. *Proceedings of the Royal Society B: Biological Sciences, 277*(1698), 3363–3372.

Friederici, A. D. (2009). Pathways to language: Fiber tracts in the human brain. *Trends in Cognitive Science, 13*(12), 175–181.

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 671–678.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1431), 459–473.

Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition, 117*(1), 1–8.

Galef, B. G., Jr. (1993). Functions of social learning about food: A causal analysis of effects of diet novelty on preference transmission. *Animal Behaviour, 46*(2), 257–265.

Galef, B. G., Jr. (1996). Social enhancement of food preferences in Norway rats: A brief review. In C. M. Heyes & B. G. Galef Jr. (Eds.), *Social learning in animals: The roots of culture* (pp. 49–64). San Diego, CA: Academic Press.

Galef, B. G., Jr., & Beck, M. (1985). Aversive and attractive marking of toxic and safe foods by Norway rats. *Behavioral and Neural Biology, 43*(3), 298–310.

Galef, B. G., Jr., & Henderson, P. W. (1972). Mother's milk: A determinant of the feeding preferences of weaning rat pups. *Journal of Comparative and Physiological Psychology, 78*(2), 213–219.

Galef, B. G., Jr., & Whiskin, E. E. (2008). Use of social information by sodium- and protein-deficient rats: Test of a prediction (Boyd & Richerson 1988). *Animal Behaviour, 75*(2), 627–630.

Galef, B. G., Jr., & Wigmore, S. W. (1983). Transfer of information concerning distant diets: A laboratory investigation of the 'information-centre' hypothesis. *Animal Behaviour, 31*(3), 748–758.

Galef, B. G., Jr., Mason, J. R., Preti, G., & Bean, N. J. (1988). Carbon disulfide: A semiochemical mediating socially-induced diet choice in rats. *Physiology & Behavior, 42*(2), 119–124.

Galef, B. G., Jr., Wigmore, S. W., & Kennett, D. J. (1983). A failure to find socially mediated taste aversion learning in Norway rats (*R. norvegicus*). *Journal of Comparative Psychology, 97*(4), 358–363.

Gallagher, S. (2007). Simulation trouble. *Social Neuroscience, 2*(3–4), 353–365.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), 593–609.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience.* Malden, MA: Wiley-Blackwell.

Gallup, G. G. (1977). Tonic immobility: The role of fear and predation. *The Psychological Record, 27*, 41–61.

Ganis, G., Schendan, H. E., & Kosslyn, S. M. (2007). Neuroimaging evidence for object model verification theory: Role of prefrontal control in visual object categorization. *NeuroImage, 34*(1), 384–398.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science, 21*(12), 1845–1853.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(3), 123–124.

Gärdenfors, P. (1996). Cued and detached representations in animal cognition. *Behavioural Processes, 35*(1–3), 263–273.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought.* Cambridge, MA: MIT Press.

Gauthier, I., & Nelson, C. A. (2001). The development of face expertise. *Current Opinion in Neurobiology, 11*(2), 219–224.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37*(12), 1673–1682.

Gavrilets, S. (1997). Evolution and speciation on holey adaptive landscapes. *Trends in Ecology and Evolution, 12*(8), 307–312.

Gazzaniga, M. S. (Ed.). (2009). *The cognitive neurosciences* (4th ed.). Cambridge, MA: MIT Press.

Gehring, W. J., & Ikeo, K. (1999). *Pax 6*: Mastering eye morphogenesis and eye evolution. *Trends in Genetics, 15*(9), 371–377.

Gelman, R., & Williams, E. M. (1998). Enabling constraints for cognitive development and learning: Domain specificity and epigenesis. In W. Damon (Ed.), *Handbook of Child Psychology: Vol. 2. Cognition, perception, and language* (5th ed., pp. 575–630). Hoboken, NJ: Wiley.

Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought. New York: Oxford University Press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7*(7), 287–292.

Gergely, G., & Csibra, G. (2006). Sylvia's recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. In N. J. Enfield & S. C. Levenson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 229–255). Oxford, UK: Berg.

Gergely, G., & Watson, J. S. (1999). Early socio-emotional development: Contingency perception and the social-biofeedback model. In P. Rochat (Ed.), *Early social cognition: Understanding others in the first months of life* (pp. 101–136). Mahwah, NJ: Lawrence Erlbaum.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature, 415*(6873), 755.

Gergely, G., Egyed, K., & Király, I. (2007). On pedagogy. *Developmental Science, 10*(1), 139–146.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193.

German, T. P., & Barrett, H. C. (2005). Functional fixedness in a technologically sparse culture. *Psychological Science, 10*(1), 1–5.

German, T. P., & Defeyter, M. A. (2000). Immunity to functional fixedness in young children. *Psychonomic Bulletin & Review, 7*(4), 707–712.

German, T. P., & Johnson, S. C. (2002). Function and the origins of the design stance. *Journal of Cognition and Development, 3*(3), 279–300.

Gibbs, R. W., Jr., & Van Orden, G. C. (2010). Adaptive cognition without massive modularity. *Language and Cognition, 2*(2), 149–176.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506–528.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gilad, Y., Man, O., Pääbo, S., & Lancet, D. (2003). Human specific loss of olfactory receptor genes. *Proceedings of the National Academy of Sciences, 100*(6), 3324–3327.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron, 31*(5), 681–697.

Gintis, H. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton, NJ: Princeton University Press.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research, 21*(2), 241–253.

Giraldeau, L.-A., Valone, T. J., & Templeton, J. J. (2002). Potential disadvantages of using socially acquired information. *Philosophical Transactions of the Royal Society B: Biological Sciences, 357*(1427), 1559–1566.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. New York: Oxford University Press.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*(5), 447–474.

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A landscape takes shape. *Cell, 128*(4), 635–638.

Goldsmith, T. H. (1990). Optimization, constraint, and history in the evolution of eyes. *Quarterly Review of Biology, 65*(3), 281–322.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.

Goodman, N. (1954). *Fact, fiction and forecast*. London: Athlone Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gordeliy, V. I., Labahn, J., Moukhametzianov, R., Efremov, R., Granzin, J., Schlesinger, R., & Engelhard, M. (2002). Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex. *Nature, 419*(6906), 484–487.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science, 306*(5695), 496–499.

Gould, J. M., & Marler, P. (1987). Learning by instinct. *Scientific American, 256*(1), 74–85.

Gould, S. J. (1977). *Ontogeny and phylogeny*. Cambridge, MA: Harvard University Press.

Gould, S. J. (1991). The disparity of the Burgess Shale arthropod fauna and the limits of cladistic analysis: Why we must strive to quantify morphospace. *Paleobiology, 17*(4), 411–423.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society B: Biological Sciences, 205*(1161), 581–598.

Grafen, A. (1984). Natural selection, kin selection and group selection. In J. R. Krebs, & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (2nd ed., pp. 62–84). Oxford, UK: Blackwell Scientific.

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences, 12*(10), 381–387.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.

Gray, R. D., Heaney, M., & Fairhall, S. (2003). Evolutionary psychology and the challenge of adaptive explanation. In K. Sterelny & J. Fitness (Eds.), *From mating to mentality: Evaluating evolutionary psychology* (pp. 247–268). New York: Psychology Press.

Green, D. M., & Swets J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11*(8), 322–323.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. *Child Development, 58*(3), 539–559.

Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Griffin, A. S., Evans, C. S., & Blumstein, D. T. (2001). Learning specificity in acquired predator recognition. *Animal Behaviour, 62*(3), 577–589.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.

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

Griffiths, P. E., & Gray, R. D. (1994). Developmental systems and evolutionary explanation. *Journal of Philosophy, 91*(6), 277–304.

Griffiths, P. E., & Gray, R. D. (2001). Darwinism and developmental systems. In S. Oyama, P. E. Griffiths, & R. D. Gray (Eds.), *Cycles of contingency: Developmental systems and evolution* (pp. 195–218). Cambridge, MA: MIT Press.

Griffiths, P., Machery, E., & Linquist, S. (2009). The vernacular concept of innateness. *Mind & Language, 24*(5), 605–630.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*(3), 441–480.

Grossman, E. D., Jardine, N. L., & Pyles J. A. (2010). fMR-adaptation reveals invariant coding of biological motion on the human STS. *Frontiers in Human Neuroscience, 4*(15), 1–18.

Guthrie, S. (1993). *Faces in the clouds: A new theory of religion*. New York: Oxford University Press.

Guzowski, J. F., Setlow, B., Wagner, E. K., & McGaugh, J. L. (2001). Experience-dependent gene expression in the rat hippocampus after spatial learning: A comparison of immediate-early genes *Arc, c-fos,* and *zif268*. *Journal of Neuroscience, 21*(14), 5089–5098.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002.

Hall, B. K. (1995). Homology and embryonic development. In M. K. Hecht, R. J. Macintyre, & M. T. Clegg (Eds.), *Evolutionary Biology* (Vol. 28, pp. 1–37). New York: Plenum Press.

Hall, B. K. (2003). Descent with modification: The unity underlying homology and homoplasy as seen through an analysis of development and evolution. *Biological Reviews, 78*(3), 409–433.

Hall, J., Thomas, K. L., & Everitt, B. J. (2000). Rapid and selective induction of *BDNF* expression in the hippocampus during contextual learning. *Nature Neuroscience, 3*(6), 533–535.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour: I. *Journal of Theoretical Biology, 7*(1), 1–16.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour: II. *Journal of Theoretical Biology, 7*(1), 17–52.

Hamilton, W. D., Axelrod, R., & Tanese, R. (1990). Sexual reproduction as an adaptation to resist parasites (A review). *Proceedings of the National Academy of Sciences, 87*(9), 3566–3573.

Hamlin, J. K., Hallinan, E. V., & Woodward, A. L. (2008). Do as I do: 7-month-old infants selectively reproduce others' goals. *Developmental Science, 11*(4), 487–494.

Hammerstein, P. (Ed.). (2003). *Genetic and cultural evolution of cooperation.* Cambridge, MA: MIT Press.

Hare, B., & Tomasello, M. (2005). Human-like social skills in dogs? *Trends in Cognitive Science, 9*(9), 439–444.

Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The domestication of social cognition in dogs. *Science, 298*(5598), 1634–1636.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*(1), 139–151.

Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review, 102*(3), 458–489.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*(1), 81–91.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10*(1), 47–66.

Hatano, G., & Inagaki, K. (2000). Domain-specific constraints of conceptual development. *International Journal of Behavioral Development, 24*(3), 267–275.

Hauser, M. D. (1997). Artifactual kinds and functional design features: What a primate understands without language. *Cognition, 64*(3), 285–308.

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), 1569–1579.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223–233.

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics, 169*(2), 1157–1164.

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology, 81*(6), 1028–1041.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Heider, F. (1958). *The psychology of interpersonal relations.* Hillsdale, NJ: Lawrence Erlbaum.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*(2), 243–259.

Heinrich, B., & Pepper, J. W. (1998). Influence of competitors on caching behaviour in the common raven, *Corvus corax*. *Animal Behaviour, 56*(5), 1083–1090.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior, 19*(4), 215–241.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79–89.

Henrich, J., & Boyd, R. (2002). On modeling cognition and culture: Why cultural evolution does not require replication of representations. *Journal of Cognition and Culture, 2*(2), 87–112.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology, 12*(3), 123–135.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Hepper, P. G. (1988). Adaptive fetal learning: Prenatal exposure to garlic affects postnatal preference. *Animal Behaviour, 36*(3), 935–936.

Hess, E. H. (1964). Imprinting in birds. *Science, 146*(3648), 1128–1139.

Heyes, C. (2009). Evolution, development and intentional control of imitation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2293–2298.

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*(1), 101–114.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience, 21*(7), 1229–1243.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1–2), 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.

Hill, K., Kaplan, H., Hawkes, K., & Hurtado, A. M. (1987). Foraging decisions among the Aché hunter-gatherers: New data and implications for optimal foraging models. *Ethology and Sociobiology, 8*(1), 1–36.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science, 30*(1), 3–41.

Holland, J. H. (1992). Complex adaptive systems. *Daedalus, 121*(1), 17–30.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Hollos, M., Leis, P. E., & Turiel, E. (1986). Social reasoning in Ijo children and adolescents in Nigerian communities. *Journal of Cross-Cultural Psychology, 17*(3), 352–374.

Horner, V., & Whiten, A. (2007). Learning from others' mistakes? Limits on understanding a trap-tube task by young chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Journal of Comparative Psychology, 121*(1), 12–21.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology, 195*(1), 215–243.

Hughes, A. L. (1997). Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. *Molecular Biology and Evolution, 14*(1), 1–5.

Hull, D. L. (1976). Are species really individuals? *Systematic Biology, 25*(2), 174–191.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). New York: Cambridge University Press.

Humphries, D. A., & Driver, P. M. (1970). Protean defence by prey animals. *Oecologia, 5*(4), 285–302.

Hunt, G. R., & Gray, R. D. (2004). The crafting of hook tools by wild New Caledonian crows. *Proceedings of the Royal Society B: Biological Sciences, 271*(Suppl. 3), S88–S90.

Hutchinson, J. M. C., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour, 75*(4), 1331–1349.

Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience, 7*(12), 942–951.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), e79.

Irons, W. (1998). Adaptively relevant environments versus the environment of evolutionary adaptedness. *Evolutionary Anthropology, 6*(6), 194-204.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Jablonka, E., & Lamb, M. J. (2005). *Evolution in four dimensions: Genetic, epigenetic, behavioral, and symbolic variation in the history of life.* Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* New York: Oxford University Press.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology, 3*(3), 318–356.

Jacob, P. (2011). The direct-perception model of empathy: A critique. *Review of Philosophy and Psychology, 2*(3), 519–540.

Jacobs, R. A. (1997). Nature, nurture, and the development of functional specializations: A computational approach. *Psychonomic Bulletin & Review, 4*(3), 299–309.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics, 33*(3s), 245–254.

Jang, K. L., Livesley, W. J., & Vemon, P. A. (1996). Heritability of the big five personality dimensions and their facets: A twin study. *Journal of Personality, 64*(3), 577–592.

Janzen, D. H., & Martin, P. S. (1982). Neotropical anachronisms: The fruits the gomphotheres ate. *Science, 215*(4528), 19–27.

Jékely, G. (2009). Evolution of phototaxis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1531), 2795–2808.

Jeschke, J. M., & Tollrian, R. (2005). Effects of predator confusion on functional responses. *Oikos, 111*(3), 547–555.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*(2), 201–211.

Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends in Cognitive Sciences, 8*(2), 71–78.

Johnson, K. L., & Shiffrar, M. (Eds.). (2013). *People watching: Social, perceptual, and neurophysiological studies of body perception.* New York: Oxford University Press.

Johnson, M. H., & Vecera, S. P. (1996). Cortical differentiation and neurocognitive development: The parcellation conjecture. *Behavioural Processes, 36*(2), 195–212.

Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1431), 549–559.

Jones, E. E., Kannouse, D. E., Kelley, H. H., Nisbett, R. E., Valins, S., & Weiner, B. (Eds.). (1972). *Attribution: Perceiving the causes of behavior.* Morristown, NJ: General Learning Press.

Joyce, R. (2006). *The evolution of morality.* Cambridge, MA: MIT Press.

Jusczyk, P. W., & Luce, P. A. (2002). Speech perception and spoken word recognition: Past and present. *Ear and Hearing, 23*(1), 2–40.

Kaas, J. H. (1984). Duplication of brain parts in evolution. *Behavioral and Brain Sciences, 7*(3), 342–343.

Kaas, J. H. (1989). The evolution of complex sensory systems in mammals. *Journal of Experimental Biology, 146*(1), 165–176.

Kaas, J. H. (2000). Why is brain size so important: Design problems and solutions as neocortex gets bigger or smaller. *Brain and Mind, 1*(1), 7–23.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus & Giroux.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175–219.

Kanazawa, S. (2004). General intelligence as a domain-specific adaptation. *Psychological Review, 111*(2), 512–523.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*(8), 759–763.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences, 107*(25), 11163–11170.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1476), 2109–2128.

Kaplan, H., Hill, K., Lancaster, J., & Hurtado, A. M. (2000). A theory of human life history evolution: Diet, intelligence, and longevity. *Evolutionary Anthropology, 9*(4), 156–185.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences, 2*(10), 389–398.

Karmiloff-Smith, A. (2009). Preaching to the converted? From constructivism to neuroconstructivism. *Child Development Perspectives, 3*(2), 99–102.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution.* New York: Oxford University Press.

Kauffman, S., & Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology, 128*(1), 11–45.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review, 88*(3), 197–227.

Keil, F. C. (1995). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 234–267). New York: Oxford University Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7*(8), 368–373.

Kelemen, D. (1999). Functions, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences, 3*(12), 461–468.

Kelemen, D., & Carey, S. (2007). The essence of artifacts: Developing the design stance. In E. Margolis & S. Laurence (Eds.), *Creations of the mind: Theories of artifacts and their representation* (pp. 212–230). Oxford, UK: Oxford University Press.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language, 22*(2), 117–131.

Kemler Nelson, D. G. (1999). Attention to functional properties in toddlers' naming and problem-solving. *Cognitive Development, 14*(1), 77–100.

Kemler Nelson, D. G., Frankenfield, A., Morris, C., & Blair, E. (2000). Young children's use of functional information to categorize artifacts: Three factors that matter. *Cognition, 77*(2), 133–168.

Kerr, N. H., & Winograd, E. (1982). Effects of contextual elaboration on face recognition. *Memory & Cognition, 10*(6), 603–609.

Ketelaar, T., & Ellis, B. J. (2000). Are evolutionary explanations unfalsifiable? Evolutionary psychology and the Lakatosian philosophy of science. *Psychological Inquiry, 11*(1), 1–21.

Keverne, E. B., & Curley, J. P. (2004). Vasopressin, oxytocin and social behaviour. *Current Opinion in Neurobiology, 14*(6), 777–783.

Khaitovich, P., Enard, W., Lachmann, M., & Pääbo, S. (2006). Evolution of primate gene expression. *Nature Reviews Genetics, 7*(9), 693–702.

Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., & Pääbo, S. (2004). Regional patterns of gene expression in human and chimpanzee brains. *Genome Research, 14*(8), 1462–1473.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation.* Cambridge, MA: MIT Press.

Kimura, M. (1983). *The neutral theory of molecular evolution.* Cambridge, UK: Cambridge University Press.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences, 104*(30), 12577–12580.

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition, 27*(4), 623–634.

Király, I., Csibra, G., & Gergely, G. (2013). Beyond rational imitation: Learning arbitrary means actions from communicative demonstrations. *Journal of Experimental Child Psychology, 116*(2), 471–486.

Király, I., Jovanovic, B., Prinz, W., Aschersleben, G., & Gergely, G. (2003). The early origins of goal attribution in infancy. *Consciousness and Cognition, 12*(4), 752–769.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), B35–B42.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics, 5*(11), 826–837.

Kline, M.A. (2014) How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences,* in press.

Knoll, A. H., & Carroll, S. B. (1999). Early animal evolution: Emerging views from comparative biology and geology. *Science, 284*(5423), 2129–2137.

Knudsen, B., & Liszkowski, U. (2012). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy, 17*(6), 672–691.

Knudsen, E. I. (2007). Fundamental components of attention. *Annual Review of Neuroscience, 30*, 57–78.

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: Comparative studies on external morphology of the primate eye. *Journal of Human Evolution, 40*(5), 419–435.

Koch, C. (1999). *Biophysics of computation: Information processing in single neurons.* New York: Oxford University Press.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In L. M. Vaina (Ed.), *Matters of intelligence: Conceptual structures in cognitive neuroscience* (pp. 115–141). Norwell, MA: Kluwer Academic Publishers.

Kohlberg, L. (1984). *Essays on Moral Development: Vol. 2. The psychology of moral development*. New York: Harper & Row.

Kokko, H., Brooks, R., McNamara, J. M., & Houston, A. I. (2002). The sexual selection continuum. *Proceedings of the Royal Society B: Biological Sciences, 269*(1498), 1331–1340.

Komarova, N. L., Niyogi, P., & Nowak, M. A. (2001). The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology, 209*(1), 43–59.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics, 39*, 309–338.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673–676.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.

Krasnow, M. M., Truxaw, D., Gaulin, S. J. C., New, J., Ozono, H., Uono, S., & Minamoto, K. (2011). Cognitive adaptations for gathering-related navigation in humans. *Evolution and Human Behavior, 32*(1), 1–12.

Krause, J., & Ruxton, G. D. (2002). *Living in groups*. New York: Oxford University Press.

Krebs, D., & Janicki, M. (2004). Biological foundations of moral norms. In M. Schaller & C. S. Crandall (Eds.), *The psychological foundations of culture* (pp. 125–148). Mahwah, NJ: Lawrence Erlbaum.

Krebs, J. R. (2009). The gourmet ape: Evolution and human food preferences. *American Journal of Clinical Nutrition, 90*(3), 707S–711S.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (2nd ed., pp. 380–402). Oxford, UK: Blackwell Scientific.

Krichevsky, A. M., King, K. S., Donahue, C. P., Khrapko, K., & Kosik, K. S. (2003). A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA, 9*(10), 1274–1281.

Krubitzer, L., & Huffman, K. J. (2000). Arealization of the neocortex in mammals: Genetic and epigenetic contributions to the phenotype. *Brain, Behavior and Evolution, 55*(6), 322–335.

Krubitzer, L., & Kahn, D. M. (2003). Nature versus nurture revisited: An old idea with a new twist. *Progress in Neurobiology, 70*(1), 33–52.

Kuenzi, F. M., & Carew, T. J. (1991). Identification of neuronal pathways mediating phototactic modulation of head-waving in *Aplysia californica*. *Behavioral and Neural Biology, 55*(3), 338–355.

Kurzban, R. (2010). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton, NJ: Princeton University Press.

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior, 32*(1), 4–14.

Laland, K. N., & Brown, G. R. (2006). Niche construction, human behavior, and the adaptive-lag hypothesis. *Evolutionary Anthropology, 15*(3), 95–104.

Laland, K. N., Odling-Smee, J., & Feldman, M. W. (2000). Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences*, *23*(1), 131–146.

Lamb, T. D., Collin, S. P., & Pugh, E. N., Jr. (2007). Evolution of the vertebrate eye: Opsins, photoreceptors, retina and eye cup. *Nature Reviews Neuroscience, 8*(12), 960–975.

Lancy, D. F. (1996). *Playing on the mother-ground: Cultural routines for children's development*. New York: Guilford.

Land, M. F., & Fernald, R. D. (1992). The evolution of eyes. *Annual Review of Neuroscience, 15*, 1–29.

Land, M. F., & Nilsson, D.-E. (2002). *Animal eyes.* New York: Oxford University Press.

Landau, B., Smith, L., & Jones, S. (1998). Object shape, object function, and object name. *Journal of Memory and Language, 38*(1), 1–27.

Langen, T. A. (1996). Social learning of a novel foraging skill by white-throated magpie-jays (*Calocitta formosa*, Corvidae): A field experiment. *Ethology, 102*(1), 157–166.

Larson, G. (1984). *In search of the Far Side.* Kansas City, MO: Andrews McMeel.

Lauder, G. V. (1990). Functional morphology and systematics: Studying functional patterns in an historical context. *Annual Review of Ecology and Systematics, 21*, 317–340.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science, 52*(2), 217–276.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184.

Lee, N., Mikesell, L., Joaquin, A. D. L., Mates, A. W., & Schumann, J. H. (2009). *The interactional instinct: The evolution and acquisition of language.* New York: Oxford University Press.

Lee, S.-H., & Wolpoff, M. H. (2003). The pattern of evolution in Pleistocene human brain size. *Paleobiology, 29*(2), 186–196.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434–1448.

Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition, 40*(3), 203–218.

Leekam, S., Baron-Cohen, S., Perrett, D., Milders, M., & Brown, S. (1997). Eye-direction detection: A dissociation between geometric and joint attention skills in autism. *British Journal of Developmental Psychology, 15*(1), 77–95.

Leekam, S., Perner, J., Healey, L., & Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. *British Journal of Developmental Psychology, 26*(4), 485–497.

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., & Jones, A. R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature, 445*(7124), 168–176.

Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology, 2*(1), 19–32.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94*(4), 412–426.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). New York: Cambridge University Press.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition, 25*(3), 265–288.

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity.* Cambridge, UK: Cambridge University Press.

Lewis, D. (1969). *Convention.* Cambridge, MA: Harvard University Press.

Lewis, D. (1999). *Papers in metaphysics and epistemology* (Vol. 2). New York: Cambridge University Press.

Lewontin, R. C. (1974). The analysis of variance and the analysis of causes. *American Journal of Human Genetics, 26*(3), 400–411.

Lickliter, R., & Honeycutt, H. (2003). Developmental dynamics: Toward a biologically plausible evolutionary psychology. *Psychological Bulletin, 129*(6), 819–835.

Liszkowski, U., & Tomasello, M. (2011). Individual differences in social, cognitive, and morphological aspects of infant pointing. *Cognitive Development, 26*(1), 16–29.

LoBue, V., Rakison, D. H., & DeLoache, J. S. (2010). Threat perception across the life span: Evidence for multiple converging pathways. *Current Directions in Psychological Science, 19*(6), 375-379.

Lorenz, K. Z. (1970). *Studies in animal and human behaviour* (Vol. 1) (R. Martin, Trans.). Cambridge, MA: Harvard University Press.

Low, J., & Wang, B. (2011). On the long road to mentalism in children's spontaneous false-belief understanding: Are we there yet? *Review of Philosophy and Psychology, 2*(3), 411–428.

Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science, 16*(8), 601–608.

Luo, Y., Kaufman, L., & Baillargeon, R. (2009). Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive Psychology, 58*(4), 441–486.

Lutz, C. A. (1988). *Unnatural emotions: Everyday sentiments on a Micronesian atoll and their challenge to Western theory.* Chicago: University of Chicago Press.

Lynch, A. (1996). *Thought contagion: How belief spreads through society: The new science of memes.* New York: Basic Books.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics, 154*(1), 459–473.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*(50), 19751–19756.

Macchi Cassia, V., Turati, C., & Simion, F. (2004). Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science, 15*(6), 379–383.

Macdonald, D. W. (1983). The ecology of carnivore social behaviour. *Nature, 301*(5899), 379–384.

Machery, E. (2007). Massive modularity and brain evolution. *Philosophy of Science, 74*(5), 825–838.

Machery, E. (2008). Massive modularity and the flexibility of human cognition. *Mind & Language, 23*(3), 263–272.

Mack, A. & Rock I. (1998). *Inattentional blindness.* Cambridge, MA: MIT Press.

MacLean, P. D. (1990). *The triune brain in evolution: Role in paleocerebral functions.* New York: Plenum Press.

Macnamara, J. (1982). *Names for things: A study of human learning.* Cambridge, MA: MIT Press.

Mahner, M., & Kary, M. (1997). What exactly are genomes, genotypes and phenotypes? And what about phenomes? *Journal of Theoretical Biology, 186*(1), 55–63.

Mangel, M., & Clark, C. W. (1988). *Dynamic modeling in behavioral ecology.* Princeton, NJ: Princeton University Press.

Marcus, G. (2004). *The birth of the mind: How a tiny number of genes creates the complexities of human thought.* New York: Basic Books.

Marcus, G. F. (2006). Cognitive architecture and descent with modification. *Cognition, 101*(2), 443–465.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science, 18*(5), 387–391.

Margolis, E., & Laurence, S. (Eds.). (2007). *Creations of the mind: Theories of artifacts and their representation*. Oxford, UK: Oxford University Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*(1), 57–77.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21*(3), 398–421.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition, 78*(1), 1–26.

Matthen, M. (2007). Defining vision: What homology thinking contributes. *Biology & Philosophy, 22*(5), 675–689.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260.

Maynard Smith, J. (1978a). Optimization theory in evolution. *Annual Review of Ecology and Systematics, 9*, 31–56.

Maynard Smith, J. (1978b). *The evolution of sex*. New York: Cambridge University Press.

Maynard Smith, J. (1982). *Evolution and the theory of games*. New York: Cambridge University Press.

Maynard Smith, J. (1991). Honest signalling: The Philip Sidney game. *Animal Behaviour, 42*(6), 1034–1035.

Maynard Smith, J. (1998). *Shaping life: Genes, embryos, and evolution*. New Haven, CT: Yale University Press.

Maynard Smith, J., & Harper, D. (2003). *Animal signals*. New York: Oxford University Press.

Maynard Smith, J., & Szathmáry, E. (1995). *The major transitions in evolution*. New York: Oxford University Press.

Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., . . . Wolpert, L. (1985). Developmental constraints and evolution: A perspective from the Mountain Lake conference on development and evolution. *Quarterly Review of Biology, 60*(3), 265–287.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution and inheritance*. Cambridge, MA: Harvard University Press.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., & Mazoyer, B. (2001). A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association, 8*(5), 401–430.

McBrearty, S., & Brooks, A. S. (2000). The revolution that wasn't: A new interpretation of the origin of modern human behavior. *Journal of Human Evolution, 39*(5), 453–563.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, 98*(20), 11832–11835.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Science, 7*(7), 293–299.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine intelligence* (Vol. 4, pp. 463–502). Edinburgh, UK: Edinburgh University Press.

McCaughey, M. (2007). *The caveman mystique: Pop-Darwinism and the debates over sex, violence, and science*. New York: Routledge.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*(11), 465–472.

McElreath, R., & Boyd, R. (2007). *Mathematical models of social evolution: A guide for the perplexed.* Chicago: University of Chicago Press.

McGhee, G. (2007). *The geometry of evolution: Adaptive landscapes and theoretical morphospaces.* New York: Cambridge University Press.

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior, 31*(5), 309–319.

McKinnon, S. (2005). *Neo-liberal genetics: The myths and moral tales of evolutionary psychology.* Chicago: Prickly Paradigm Press.

McShea, D. W., & Brandon, R. N. (2010). *Biology's first law: The tendency for diversity and complexity to increase in evolutionary systems.* Chicago: University of Chicago Press.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology, 24*(4), 470–476.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*(5), 838–850.

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. Chicago: University of Chicago Press.

Mesoudi, A., & O'Brien, M. J. (2008). The learning and transmission of hierarchical cultural recipes. *Biological Theory, 3*(1), 63–72.

Michotte, A. (1963). *The perception of causality.* New York: Basic Books.

Midkiff, E. E., & Bernstein, I. L. (1985). Targets of learned food aversions in humans. *Physiology & Behavior, 34*(5), 839–841.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Miklósi, Á. (2007). *Dog behaviour, evolution, and cognition.* New York: Oxford University Press.

Miklósi, Á., Kubinyi, E., Topál, J., Gácsi, M., Virányi, Z., & Csányi, V. (2003). A simple reason for a big difference: Wolves do not look back at humans, but dogs do. *Current Biology, 13*(9), 763–766.

Miller, G. (2009). *Spent: Sex, evolution, and consumer behavior.* New York: Viking.

Miller, G. F., & Todd, P. M. (1995). The role of mate choice in biocomputation: Sexual selection as a process of search, optimization, and diversification. In W. Banzhaf & F. H. Eeckman (Eds.), *Evolution and biocomputation: Computational models of evolution* (pp. 169–204). Berlin: Springer-Verlag.

Miller, M. B., & Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Review of Microbiology, 55*, 165–199.

Milton, K. (1999). Nutritional characteristics of wild primate foods: Do the diets of our closest living relatives have lessons for us? *Nutrition, 15*(6), 488–498.

Mineka, S., Davidson, M., Cook, M., & Keir, R. (1984). Observational conditioning of snake fear in rhesus monkey. *Journal of Abnormal Psychology, 93*(4), 355–372.

Minsky, M. (1988). *Society of mind.* New York: Simon & Schuster.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414–417.

Mitani, J. C., & Watts, D. P. (2001). Why do chimpanzees hunt and share meat? *Animal Behaviour, 61*(5), 915–924.

Mitchell, D. R. (2007). The evolution of eukaryotic cilia and flagella as motile and sensory organelles. In G. Jékely (Ed.), *Eukaryotic membranes and cytoskeleton: Origins and evolution*. (pp. 130–140). New York: Springer.

Mithen, S. J. (1996). *The prehistory of the mind: A search for the origins of art, religion and science*. London: Thames and Hudson.

Miyashita-Lin, E. M., Hevner, R., Wassarman, K. M., Martinez, S., & Rubenstein, J. L. R. (1999). Early neocortical regionalization in the absence of thalamic innervation. *Science, 285*(5429), 906–909.

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding, 104*(2–3), 90–126.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: A two-process theory of infant face recognition. *Psychological Review, 98*(2), 164–181.

Moya, C. (2013). Evolved priors for ethnolinguistic categorization: A case study from the Quechua-Aymara boundary in the Peruvian Altiplano. *Evolution and Human Behavior, 34*(4), 265–272.

Munday, P. L., Buston, P. M., & Warner, R. R. (2006). Diversity and flexibility of sex-change strategies in animals. *Trends in Ecology and Evolution, 21*(2), 89–95.

Nakagawa, Y., Johnson, J. E., & O'Leary, D. D. M. (1999). Graded and areal expression patterns of regulatory genes and cadherins in embryonic neocortex independent of thalamocortical input. *Journal of Neuroscience, 19*(24), 10877–10885.

Namihira, M., Kohyama, J., Abematsu, M., & Nakashima, K. (2008). Epigenetic mechanisms regulating fate specification of neural stem cells. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1500), 2099–2109.

Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition, 47*(2), 121–148.

Needham, A., & Baillargeon, R. (2000). Infants' use of featural and experiential information in segregating and individuating objects: A reply to Xu, Carey and Welch (2000). *Cognition, 74*(3), 255–284.

New, J., Cosmides, L., & Tooby, J. (2007a). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences, 104*(42), 16598–16603.

New, J., Krasnow, M. M., Truxaw, D., & Gaulin, S. J. C. (2007b). Spatial adaptations for plant foraging: Women excel and calories count. *Proceedings of the Royal Society B: Biological Sciences, 274*(1626), 2679–2684.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28*(1), 28–38.

Ng, S. Y. M., Chaban, B., & Jarrell, K. F. (2006). Archaeal flagella, bacterial flagella and type IV pili: A comparison of genes and posttranslational modifications. *Journal of Molecular Microbiology and Biotechnology, 11*(3–5), 167–191.

Nichols, S. (2004a). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.

Nichols, S. (2004b). The folk psychology of free will: Fits and starts. *Mind & Language, 19*(5), 473–502.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. New York: Oxford University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Noble, J., Todd, P. M., & Tuci, E. (2001). Explaining social learning of food preferences without aversions: An evolutionary simulation model of Norway rats. *Proceedings of the Royal Society B: Biological Sciences, 268*(1463), 141–149.

Novembre, J., & Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics, 10*(11), 745–755.

Nowak, M. A. (2006). *Evolutionary dynamics: Exploring the equations of life*. Cambridge, MA: Harvard University Press.

Nowak, M. A., Plotkin, J. B., & Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature, 404*(6777), 495–498.

Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49*(2), 400–407.

Nylin, S., & Gotthard, K. (1998). Plasticity in life-history traits. *Annual Review of Entomology, 43*, 63–83.

O'Brien, R. M., & Granner, D. K. (1996). Regulation of gene expression by insulin. *Physiology Review, 76*(4), 1109–1161.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659–677.

Ohala, J. J., Hinton, L., & Nichols, J. (Eds.). (1994). *Sound symbolism*. New York: Cambridge University Press.

Öhman, A. (2005). The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology, 30*(10), 953–958.

Öhman, A. (2006). Making sense of emotion: Evolution, reason & the brain. *Daedalus, 135*(3), 33–45.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108*(3), 483–522.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*(3), 466–478.

Öhman, A., Fredrikson, M., Hugdahl, K., & Rimmö, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General, 105*(4), 313–337.

Ohno, S. (1970). *Evolution by gene duplication*. New York: Springer-Verlag.

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences, 99*(25), 16134–16137.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics, 10*(8), 531–539.

Ottoni, E. B., Dogo de Resende, B., & Izar, P. (2005). Watching the best nutcrackers: What capuchin monkeys (*Cebus apella*) know about others' tool-using skills. *Animal Cognition, 8*(4), 215–219.

Oyama, S. (2000). *The ontogeny of information: Developmental systems and evolution* (2nd ed.). Durham, NC: Duke University Press.

Oyama, S., Griffiths, P. E., & Gray, R. D. (Eds.). (2001). *Cycles of contingency: Developmental systems and evolution*. Cambridge, MA: MIT Press.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Panksepp, J., & Panksepp, J. B. (2000). The seven sins of evolutionary psychology. *Evolution and Cognition, 6*(2), 108–131.

Papoušek, M., Bornstein, M. H., Nuzzo. C., Papoušek. H., & Symmes, D. (1990). Infant responses to prototypical melodic contours in parental speech. *Infant Behavior and Development, 13*(4), 539–545.

Pascalis, O., & Bachevalier, J. (1998). Face recognition in primates: A cross-species study. *Behavioural Processes, 43*(1), 87–96.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance, 10*(3), 358–377.

Pavlov, I. P. (1902). *The work of the digestive glands* (W. H. Thompson, Trans.). Philadelphia: Griffin.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 534–552.

Penfield, W., & Jasper, H. (1954). *Epilepsy and the functional anatomy of the human brain*. Boston: Little, Brown and Co.

Penfield, W., & Rasmussen, T. (1950). *The cerebral cortex of man: A clinical study of localization of function*. New York: MacMillan.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind.' *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 731–744.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*(2), 109–130.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308*(5719), 214–216.

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development, 60*(3), 689–700.

Perreault, C. (2012). The pace of cultural evolution. *PLOS ONE, 7*(9), e45150.

Perreault, C., Moya, C., & Boyd, R. (2012). A Bayesian approach to the evolution of social learning. *Evolution and Human Behavior, 33*(5), 449–459.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics, 39*(10), 1256–1260.

Perry, G., & Pianka, E. R. (1997). Animal foraging: Past, present and future. *Trends in Ecology and Evolution, 12*(9), 360–364.

Perry, S. (2011). Social traditions and social learning in capuchin monkeys (*Cebus*). *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), 988–996.

Perry, S., & Manson, J. (2008). *Manipulative monkeys: The capuchins of Lomas Barbudal*. Cambridge, MA: Harvard University Press.

Peterson, M. (2009). *An introduction to decision theory.* New York: Cambridge University Press.

Petroski, H. (1992). *The evolution of useful things: How everyday artifacts—from forks and pins to paper clips and zippers—came to be as they are.* New York: Knopf.

Philippi, T., & Seger, J. (1989). Hedging one's evolutionary bets, revisited. *Trends in Ecology and Evolution, 4*(2), 41–44.

Piaget, J. (1954). *The construction of reality in the child.* London: Routledge & Kegan Paul.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science, 306*(5695), 499–503.

Pinker, S. (1979). Formal models of language learning. *Cognition, 7*(3), 217–283.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1997). *How the mind works.* New York: Norton.

Pinker, S. (2005). So how *does* the mind work? *Mind & Language, 20*(1), 1–24.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* New York: Viking.

Pinker, S. (2011). *Words and rules: The ingredients of language.* New York: HarperCollins.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*(4), 707–727.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition, 95*(2), 201–236.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*(1–2), 73–193.

Pinker, S., & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Science, 6*(6), 456–463.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643–675.

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science, 20*(11), 1364–1372.

Popper, K. (1959). *The logic of scientific discovery.* London: Hutchinson.

Posner, M. I., & Raichle, M. E. (1998). The neuroimaging of human brain function. *Proceedings of the National Academy of Sciences, 95*(3), 763–764.

Povinelli, D. J., & Eddy, T. J. (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development, 61*(2, Serial No. 247).

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Science, 7*(4), 157–160.

Prasada, S., Ferenz, K., & Haskell, T. (2002). Conceiving of entities as objects and stuff. *Cognition, 83*(2), 141–165.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioural and Brain Science, 1*(4), 515–526.

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. *Journal of Anatomy, 197*(3), 335–359.

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology, 22*(3–4), 262–275.

Profet, M. (1988). The evolution of pregnancy sickness as protection to the embryo against Pleistocene teratogens. *Evolutionary Theory, 8*(3), 177–190.

Prud'homme, B., Gompel, N., & Carroll, S. B. (2007). Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences, 104*(Suppl. 1), 8605–8612.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*(1–2), 127–158.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences, 20*(4), 537–556.

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Quine, W. V. O. (1969). *Ontological relativity and other essays.* New York: Columbia University Press.

Raff, R. A. (1996). *The shape of life: Genes, development, and the evolution of animal form.* Chicago: University of Chicago Press.

Raff, R. A. (2000). Evo-devo: The evolution of a new discipline. *Nature Reviews Genetics, 1*(1), 74–79.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118*(1), 57–75.

Rakoczy, H. (2008). Taking fiction seriously: Young children understand the normative structure of joint pretence games. *Developmental Psychology, 44*(4), 1195–1201.

Rakoczy, H., Brosche, N., Warneken, F., & Tomasello, M. (2009). Young children's understanding of the context-relativity of normative rules in conventional games. *British Journal of Developmental Psychology, 27*(2), 445–456.

Rasskin-Gutman, D. (2005). Modularity: Jumping forms within morphospace. In W. Callebaut & D. Rasskin-Gutman (Eds.), *Modularity: Understanding the development and evolution of natural complex systems* (pp. 207–220). Cambridge, MA: MIT Press.

Rastogi, S., & Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology, 5*(1), 28.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience, 12*(6), 718–724.

Read, D. (2001). Is time-discounting hyperbolic or subadditive? *Journal of Risk and Uncertainty, 23*(1), 5–32.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., & Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature, 468*(7327), 1053–1060.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., & Laland, K. N. (2010). Why copy others? Insights from the social learning strategies tournament. *Science, 328*(5975), 208–213.

Riba, J., Valle, M., Urbano, G., Yritia, M., Morte, A., & Barbanoj, M. J. (2003). Human pharmacology of ayahuasca: Subjective and cardiovascular effects, monoamine metabolite excretion, and pharmacokinetics. *Journal of Pharmacology and Experimental Therapeutics, 306*(1), 73–83.

Rice, S. H. (2002). A general population genetic theory for the evolution of developmental interactions. *Proceedings of the National Academy of Sciences, 99*(24), 15518–15523.

Rice, W. R. (1996). Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature, 381*(6579), 232–234.

Richardson, R. C. (2007). *Evolutionary psychology as maladapted psychology.* Cambridge, MA: MIT Press.

Richerson, P. J., & Boyd, R. (2001). The evolution of subjective commitment to groups: A tribal instincts hypothesis. In R. M. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 186–220). New York: Russell Sage Foundation.

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.

Richmond, B. G., & Jungers, W. L. (2008). *Orrorin* tugenensis femoral morphology and the evolution of hominin bipedalism. *Science, 319*(5870), 1662–1665.

Richter, C. P., Holt, L. E., Jr., & Barelare, B., Jr. (1937). Vitamin B1 craving in rats. *Science, 86*(2233), 354–355.

Ridley, M. (1995). *The Red Queen: Sex and the evolution of human nature*. New York: Penguin Books.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*(2), 162–168.

Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., & Behrens, T. E. J. (2008). The evolution of the arcuate fasciculus revealed with comparative DTI. *Nature Neuroscience, 11*(4), 426–428.

Ristau, C. A. (1991). Aspects of the cognitive ethology of an injury-feigning bird, the Piping Plover. In C. A. Ristau (Ed.), *Cognitive ethology: The minds of other animals* (pp. 91–126). Hillsdale, NJ: Lawrence Erlbaum.

Ritter, R. C. (2004). Gastrointestinal mechanisms of satiation for food. *Physiology & Behavior, 81*(2), 249–273.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192.

Rodin, J., Mancuso, J., Granger, J., & Nelbach, E. (1991). Food cravings in relation to body mass index, restraint and estradiol levels: A repeated measures study in healthy women. *Appetite, 17*(3), 177–185.

Roland, P. E., & Zilles, K. (1998). Structural divisions and functional fields in the human cerebral cortex. *Brain Research Reviews, 26*(2–3), 87–105.

Rosaldo, M. Z. (1982). The things we do with words: Ilongot speech acts and speech act theory in philosophy. *Language in Society, 11*(2), 203–237.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.

Rose, H., & Rose, S. (Eds.). (2000). *Alas, poor Darwin: Arguments against evolutionary psychology*. New York: Random House.

Rothstein, S. I. (1990). A model system for coevolution: Avian brood parasitism. *Annual Review of Ecology and Systematics, 21*, 481–508.

Rozin, P., & Fallon, A. E. (1987). A perspective on disgust. *Psychological Review, 94*(1), 23–41.

Rozin, P., & Kalat, J. W. (1971). Specific hungers and poison avoidance as adaptive specializations of learning. *Psychological Review, 78*(6), 459–486.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology, 50*(4), 703–712.

Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology, 30*(1), 87–104.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986–1987). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vols. 1–2). Cambridge, MA: MIT Press.

Rutherford, M. D., & Kuhlmeier, V. A. (Eds.). (2013). *Social perception: Detection and interpretation of animacy, agency, and intention.* Cambridge, MA: MIT Press.

Ruvinksy, I., & Gibson-Brown, J. J. (2000). Genetic and developmental bases of serial homology in vertebrate limb evolution. *Development, 127*(24), 5233–5244.

Saad, G. (2007). *The evolutionary bases of consumption.* Mahwah, NJ: Lawrence Erlbaum.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., & The International HapMap Consortium. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature, 449*(7164), 913–918.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current Directions in Psychological Science, 12*(4), 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Sahlins, M. D. (2008). *The Western illusion of human nature: With reflections on the long history of hierarchy, equality and the sublimation of anarchy in the West, and comparative notes on other conceptions of the human condition.* Chicago: Prickly Paradigm.

Salvini-Plawen, L. V., & Mayr, E. (1977). On the evolution of photoreceptors and eyes. In M. K. Hecht, W. C. Steere, & B. Wallace (Eds.), *Evolutionary Biology* (Vol. 10, pp. 207–263). New York: Plenum Press.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *British Journal for the Philosophy of Science, 49*(4), 575–602.

Samuels, R. (2002). Nativism in cognitive science. *Mind & Language, 17*(3), 233–265.

Samuels, R. (2004). Innateness in cognitive science. *Trends in Cognitive Sciences, 8*(3), 136–141.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition, 108*(3), 662–674.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*(2), 235–239.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Scheibehenne, B., Todd, P. M., & Wansink, B. (2010). Dining in the dark: The importance of visual cues for food consumption and satiety. *Appetite, 55*(3), 710–713.

Scher, S. J., & Rauscher, F. (Eds.). (2002). *Evolutionary psychology: Alternative approaches.* Norwell, MA: Kluwer.

Schlichting, C. D., & Pigliucci, M. (1995). Gene regulation, quantitative genetics and the evolution of reaction norms. *Evolutionary Ecology, 9*(2), 154–168.

Schlichting, C. D., & Pigliucci, M. (1998). *Phenotypic evolution: A reaction norm perspective.* Sunderland, MA: Sinauer.

Schoenemann, P. T., Sheehan, M. J., & Glotzer, L. D. (2005). Prefrontal white matter volume is disproportionately larger in humans than in other primates. *Nature Neuroscience, 8*(2), 242–252.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*(8), 299–309.

Searle, J. R. (1995). *The construction of social reality.* New York: Free Press.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515–530.

Seeley, T. D. (2009). *The wisdom of the hive: The social physiology of honey bee colonies.* Cambridge, MA: Harvard University Press.

Selfridge, O. G., & Neisser, U. (1960). Pattern recognition by machine. *Scientific American, 203*(3), 60–68.

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77*(5), 406–418.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*(35), 15073–15078.

Semple, S., & McComb, K. (1996). Behavioural deception. *Trends in Ecology and Evolution, 11*(10), 434–437.

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences, 110*(40), 15937–15942.

Shallice, T. (1988). *From neuropsychology to mental structure.* New York: Cambridge University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana, IL: University of Illinois Press.

Shepard, R. N. (1992). The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 495–532). New York: Oxford University Press.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1*(1), 2–28.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review, 94*(4), 439–454.

Shi, Y., & Yokoyama, S. (2003). Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proceedings of the National Academy of Sciences, 100*(14), 8308–8313.

Shostak, M. (1981). *Nisa: The life and words of a !Kung woman.* Cambridge, MA: Harvard University Press.

Shubin, N. (2008). *Your inner fish: A journey into the 3.5-billion-year history of the human body.* New York: Pantheon Books.

Shutts, K., Kinzler, K. D., McKee, C. B., & Spelke, E. S. (2009). Social information guides infants' selection of foods. *Journal of Cognition and Development, 10*(1–2), 1–17.

Silk, J. B. (2007). The adaptive value of sociality in mammalian groups. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 539–559.

Simon, H. A. (1979). *Models of thought* (Vol. 1). New Haven, CT: Yale University Press.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–19.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261–267.

Simpson, G. G. (1953). The Baldwin effect. *Evolution, 7*(2), 110–117.

Sinclair, A. H. (1998). Human sex determination. *Journal of Experimental Zoology, 281*(5), 501–505.

Slatkin, M. (1974). Hedging one's evolutionary bets. *Nature, 250*(5469), 704–705.

Slaughter, V., & Corbett, D. (2007). Differential copying of human and nonhuman models at 12 and 18 months of age. *European Journal of Developmental Psychology, 4*(1), 31–45.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.

Smetana, J. D. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52*(4), 1333–1336.

Smith, C. C., & Reichman, O. J. (1984). The evolution of food caching by birds and mammals. *Annual Review of Ecology and Systematics, 15*, 329–351.

Smith, E. A. (1983). Anthropological applications of optimal foraging theory: A critical review. *Current Anthropology, 24*(5), 625–640.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22*(4), 377–386.

Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*(8), 343–348.

Sniegowski, P. D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of *E. coli*. *Nature, 387*(6634), 703–705.

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition, 95*(1), 1–30.

Song, H.-J., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology, 44*(6), 1789–1795.

Southgate, V. (2013). Early manifestations of mindreading. In S. Baron-Cohen, H. Tager-Flusberg, & M. V. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 3–18). New York: Oxford University Press.

Southgate, V., Csibra, G., Kaufman, J., & Johnson, M. H. (2008). Distinct processing of objects and faces in the infant brain. *Journal of Cognitive Neuroscience, 20*(4), 741–749.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*(1), 29–56.

Spelke, E. S., & Kinzler, K. D. (2009). Innateness, learning, and rationality. *Child Development Perspectives, 3*(2), 96–98.

Spencer, J. P., Blumberg, M. S., McMurray, B., Robinson, S. R., Samuelson, L. K., & Tomblin, J. B. (2009). Short arms and talking eggs: Why we should no longer abide the nativist-empiricist debate. *Child Development Perspectives, 3*(2), 79–87.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). New York: Cambridge University Press.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Malden, MA: Blackwell.

Sperber, D. (2001). In defense of massive modularity. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in honor of Jacques Mehler* (pp. 47–57). Cambridge, MA: MIT Press.

Sperber, D. (2007). Seedless grapes: Nature and culture. In E. Margolis & S. Laurence (Eds.), *Creations of the mind: Theories of artifacts and their representation* (pp. 124–137). Oxford, UK: Oxford University Press.

Sperber, D., & Hirschfeld, L. A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences, 8*(1), 40–46.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, UK: Blackwell.

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron, 35*(6), 1157–1165.

Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences, 8*(9), 418–425.

Sripada, S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Vol. 2. Culture and cognition* (pp. 280–301). New York: Oxford University Press.

Stanford, C. B., & Bunn, H. T. (Eds.). (2001). *Meat-eating and human evolution.* New York: Oxford University Press.

Stankowich, T., & Blumstein, D. T. (2005). Fear in animals: A meta-analysis and review of risk assessment. *Proceedings of the Royal Society B: Biological Sciences, 272*(1581), 2627–2634.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin.* Chicago: University of Chicago Press.

Stearns, S. C. (1989). The evolutionary significance of phenotypic plasticity. *BioScience, 39*(7), 436–445.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory.* Princeton, NJ: Princeton University Press.

Stepniewska, I., Friedman, R. M., Gharbawie, O. A., Cerkevich, C. M., Roe, A. W., & Kaas, J. H. (2011). Optical imaging in galagos reveals parietal-frontal circuits underlying motor behavior. *Proceedings of the National Academy of Sciences, 108*(37), E725–E732.

Striedter, G. F. (2005). *Principles of brain evolution.* Sunderland, MA: Sinauer.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences, 30*(3), 299–313.

Sur, M., & Leamey, C. A. (2001). Development and plasticity of cortical areas and networks. *Nature Reviews Neuroscience, 2*(4), 251–262.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586.

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., & Ostrander, E. A. (2007). A single *IGF1* allele is a major determinant of small size in dogs. *Science, 316*(5821), 112.

Symons, D. (1990). Adaptiveness and adaptation. *Ethology and Sociobiology, 11*(4–5), 427–444.

Taddei, F., Radman, M., Maynard Smith, J., Toupance, B., Gouyon, P. H., & Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature, 387*(6634), 700–702.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex, 13*(1), 90–99.

Taylor, P. D., & Jonker, L. B. (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences, 40*(1–2), 145–156.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285.

Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2405–2415.

Terakita, A. (2005). The opsins. *Genome Biology, 6*(3), 213.

Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory.* Cambridge, MA: MIT Press.

Thompson, J. C., Clarke, M., Stewart, T., & Puce, A. (2005). Configural processing of biological motion in human superior temporal sulcus. *Journal of Neuroscience, 25*(39), 9059–9066.

Thornton, A., & Raihani, N. J. (2008). The evolution of teaching. *Animal Behaviour, 75*(6), 1823–1836.

Thorpe, S., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*(5), 869–876.

Todd, P. M., & Gigerenzer, G. (2001). Shepard's mirrors or Simon's scissors? *Behavioral and Brain Sciences, 24*(4), 704–705.

Tomasello, M. (1999). *The cultural origins of human cognition.* Cambridge, MA: Harvard University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tomasello, M., & Barton, M. (1994). Learning words in nonostensive contexts. *Journal of Developmental Psychology, 30*(5), 639–650.

Tomasello, M., & Call, J. (1997). *Primate cognition.* Oxford, UK: Oxford University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–691.

Tooby, J. (1982). Pathogens, polymorphism, and the evolution of sex. *Journal of Theoretical Biology, 97*(4), 557–576.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology, 11*(4–5), 375–424.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

Tooby, J., & DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In W. G. Kinzey (Ed.), *The evolution of human behavior: Primate models* (pp. 183–237). Albany, NY: State University of New York Press.

Tooby, J., Cosmides, L., & Barrett, H. C. (2003). The second law of thermodynamics is the first law of psychology: Evolutionary developmental psychology and the theory of tandem, coordinated inheritances: Comment on Lickliter and Honeycutt (2003). *Psychological Bulletin, 129*(6), 858–865.

Tooby, J., Cosmides, L., & Barrett, H. C. (2005). Resolving the debate on innate ideas: Learnability constraints and the evolved interpenetration of motivational and conceptual functions. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Vol. 1. Structure and contents* (pp. 305–337). New York: Oxford University Press.

Torday, J. S., & Rehan, V. K. (2012). *Evolutionary biology, cell-cell communication, and complex disease.* Hoboken, NJ: Wiley-Blackwell.

Toth, A. L., & Robinson, G. E. (2007). Evo-devo and the evolution of social behavior. *Trends in Genetics, 23*(7), 334–341.

Towers, S. R., & Coss, R. G. (1990). Confronting snakes in the burrow: Snake-species discrimination and antisnake tactics of two California ground squirrel populations. *Ethology, 84*(3), 177–192.

Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy, 15*(4), 434–444.

Treisman, A. M., & Gelade, G., (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry, 42*(1), 3–48.

Truxaw, D., Krasnow, M. M., Woods, C., & German, T. P. (2006). Conditions under which function information attenuates name extension via shape. *Psychological Science, 17*(5), 367–371.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*(4940), 301–306.

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge, UK: Cambridge University Press.

Udell, M. A., Dorey, N. R., & Wynne, C. D. (2008). Wolves outperform dogs in following human social cues. *Animal Behaviour, 76*(6), 1767–1773.

Uhlhaas, P. J., Roux, F., Rodriguez, E., Rotarska-Jagiela, A., & Singer, W. (2010). Neural synchrony and the development of cortical networks. *Trends in Cognitive Sciences, 14*(12), 72–80.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11*(2), 58–64.

Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology, 4*(2), 157–165.

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology, 44*(5), 1334–1338.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

Van Essen, D. C., & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron, 13*(1), 1–10.

Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy, 92*(7), 345–381.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory, 1*(1), 1–30.

Via, S., Gomulkiewicz, R., De Jong, G., Scheiner, S. M., Schlichting, C. D., & Van Tienderen, P. H. (1995). Adaptive phenotypic plasticity: Consensus and controversy. *Trends in Ecology and Evolution, 10*(5), 212–217.

Virányi, Z., Topál, J., Miklósi, Á., & Csányi, V. (2006). A nonverbal test of knowledge attribution: A comparative study on dogs and children. *Animal Cognition, 9*(1), 13–26.

Visalberghi, E., & Fragaszy, D. (2002). "Do monkeys ape?" Ten years after. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in animals and artifacts* (pp. 471–500). Cambridge, MA: MIT Press.

Visalberghi, E., & Limongelli, L. (1994). Lack of comprehension of cause-effect relations in tool-using capuchin monkeys (*Cebus apella*). *Journal of Comparitive Psychology, 108*(1), 15–22.

Vogel, S. (1988). *Life's devices: The physical world of animals and plants.* Princeton, NJ: Princeton University Press.

Waddington, C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology.* London: George Allen & Unwin.

Wagner, G. P., & Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. *Evolution, 50*(3), 967–976.

Wagner, G. P., Pavlicev, M., & Cheverud, J. M. (2007). The road to modularity. *Nature Reviews Genetics, 8*(12), 921-931.

Want, S. C., & Harris, P. L. (2001). Learning from other people's mistakes: Causal understanding in learning to use a tool. *Child Development, 72*(2), 431–443.

Want, S. C., & Harris, P. L. (2002). How do children ape? Applying concepts from the study of non-human primates to the developmental study of 'imitation' in children. *Developmental Science, 5*(1), 1–14.

Warner, R. R. (1984). Mating behavior and hermaphroditism in coral reef fishes. *American Scientist, 72*(2), 128–136.

Watson, J. S. (1985). Contingency perception in early social development. In T. M. Field & N. A. Fox (Eds.), *Social perception in infants* (pp. 157–176). Norwood, NJ: Ablex.

Wayne, R. K., & Ostrander E. A. (2007). Lessons learned from the dog genome. *Trends in Genetics, 23*(11), 557–567.

Weber, B. H., & Depew, D. J. (Eds.). (2003). *Evolution and learning: The Baldwin effect reconsidered.* Cambridge, MA: MIT Press.

Wegner, D. M. (2002). *The illusion of conscious will.* Cambridge, MA: MIT Press.

Weir, A. A. S., Chappell, J., & Kacelnik, A. (2002). Shaping of hooks in New Caledonian crows. *Science, 297*(5583), 981.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., & Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences, 95*(1), 334–339.

West-Eberhard, M. J. (2003). *Developmental plasticity and evolution.* New York: Oxford University Press.

West, A. E., Chen, W. G., Dalva, M. B., Dolmetsch, R. E., Kornhauser, J. M., Shaywitz, A. J., & Greenberg, M. E. (2001). Calcium regulation of neuronal gene expression. *Proceedings of the National Academy of Sciences, 98*(20), 11024–11031.

Wheeler, Q. D., & Meier, R. (Eds.). (2000). *Species concepts and phylogenetic theory: A debate.* New York: Columbia University Press.

White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society B: Biological Sciences, 314*(1165), 1–340.

Whitehead, H. (2000). *Food rules: Hunting, sharing, and tabooing game in Papua New Guinea.* Ann Arbor, MI: University of Michigan Press.

Whiten, A. (1996). When does smart behaviour-reading become mind-reading? In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 277–292). New York: Cambridge University Press.

Whiten, A. (2011). The scope of culture in chimpanzees, humans and ancestral apes. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), 997–1007.

Whiten, A., Horner, V., Litchfield, C. A., & Marshall-Pescini, S. (2004). How do apes ape? *Learning & Behavior, 32*(1), 36–52.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2417–2428.

Whitfield, C. W., Czikol, A.-M., & Robinson, G. E. (2003). Gene expression profiles in the brain predict behavior in individual honey bees. *Science, 302*(5643), 296–299.

Wilke, A., & Barrett, H. C. (2009). The hot hand phenomenon as a cognitive adaptation to clumped resources. *Evolution and Human Behavior, 30*(3), 161–169.

Wilke, A., Hutchinson, J. M. C., Todd, P. M., & Czienskowski, U. (2009). Fishing for the right words: Decision rules for human foraging behavior in internal search tasks. *Cognitive Science, 33*(3), 497–529.

Williamson, R. A., Meltzoff, A. N., & Markman, E. M. (2008). Prior experiences and perceived efficacy influence 3-year-olds' imitation. *Developmental Psychology, 44*(1), 275–285.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.

Wimsatt, W. C., & Schank, J. C. (2004). Generative entrenchment, modularity and evolvability: When genic selection meets the whole organism. In G. Schlosser & G. Wagner (Eds.), *Modularity in development and evolution* (pp. 359–394). Chicago: University of Chicago Press.

Wood, B., & Collard, M. (1999). The human genus. *Science, 284*(5411), 65–71.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34.

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development, 22*(2), 145–160.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11*(1), 73–77.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In D. F. Jones, *Proceedings of the Sixth International Congress of Genetics* (Vol. 1, pp. 356–366). Austin, TX: Genetics Society of America.

Wurtman, R. J., & Wurtman, J. J. (1986). Carbohydrate craving, obesity and brain serotonin. *Appetite, 7*(Suppl.), 99–103.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358*(6389), 749–750.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition, 85*(3), 223–250.

Xu, F., & Carey, S. (2000). The emergence of kind concepts: A rejoinder to Needham & Baillargeon (2000). *Cognition, 74*(3), 285–301.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272.

Xu, F., Carey, S., & Welch, J. (1999). Infants' ability to use object kind information for object individuation. *Cognition, 70*(2), 137–166.

Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396–1405.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235–8240.

Young, R. L., & Wagner, G. P. (2011). Why ontogenetic homology criteria can be misleading: Lessons from digit identity transformations. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 316B*(3), 165–170.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*(6), 979–1008.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*(1), 3–21.

Zahavi, D. (2011). Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology, 2*(3), 541–558.

Zakon, H. H. (2012). Adaptive evolution of voltage-gated sodium channels: The first 800 million years. *Proceedings of the National Academy of Sciences, 109*(Suppl. 1), 10619–10625.

Zellner, D. A., Garriga-Trillo, A., Rohm, E., Centeno, S., & Parker, S. (1999). Food liking and craving: A cross-cultural approach. *Appetite, 33*(1), 61–70.

Zeng, Y. (2009). Regulation of the mammalian nervous system by microRNAs. *Molecular Pharmacology, 75*(2), 259–264.

Zentall, T., & Galef, B. G., Jr. (Eds.). (1988). *Social learning: Psychological and biological perspectives*. Hillsdale, NJ: Lawrence Erlbaum.

# INDEX


aboutness 58
adaptation 2–12, 23–31, 34, 39, 41, 52–54, 57–58, 62, 77–79, 95, 100, 103, 104, 155, 157–158, 160, 164, 166–169, 176, 180, 190, 192–194, 196, 198, 202–204, 209–215, 217–222, 225, 238, 240–249, 251–252, 256–257, 262, 266–267, 269, 273, 287–288, 290, 292, 300, 302, 304, 308, 324, 333–338
  brain 5–7, 333. *See also* adaptation (mental)
  cognitive 211. *See also* adaptation (mental)
  mental 4, 11, 57, 155, 215, 217, 252, 266, 287, 292, 338
  psychological 5, 78, 211, 269. *See also* adaptation (mental)
adaptive lag 204. *See* evolutionary disequilibrium
adaptive landscape 50–55, 60, 64n3, 76, 150, 162–163, 200–202, 214, 256, 291, 330. *See* fitness landscape
adaptive radiation 4, 301. *See also* niche differentiation
adaptive target 19, 30, 35, 191, 195–199
additivism 17, 19, 30–32, 63, 107, 181, 202, 237, 241. *See also* subtractivism
affordance 66, 99–100, 249–251
agency 99, 109, 126, 146, 309. *See* intentional agency
allegory of the cave 126, 265
altruism 142, 220
ancestrality 6, 77, 119, 169, 171, 193–194, 196–197, 233, 266, 286–287, 300, 307–310, 325–327
  aspectual 194
anger 255, 275, 278
animate monitoring hypothesis 113
antagonistic selection 173, 222
arms race 210, 222. *See also* Red Queen
artifact 12, 26, 53–54, 114, 198, 217, 225, 243–252, 327–328, 337
assembly line system 269–274, 280, 284, 286, 294
assumption 39, 87–91, 95, 98, 129, 139, 156, 162, 172, 190, 218–221, 225, 230, 271, 323, 334–335. *See also* prior, ontological commitment
  equipotentiality 91, 235
  genericity 240, 251
  immutability 321
  rationality 122
  whole-object 39, 239
attention 44, 76, 110, 116, 135–136, 143–144, 201–202, 221, 231, 235, 239–240, 271, 293–294, 310
  allocation of 112–113, 284–286
  spotlight model of 112
  shared 240. *See also* triadic awareness
attractor 235
autapomorphy 46, 291, 325. *See also* trait (derived)
automaticity 5–8, 68, 89, 257, 264–271, 276, 279, 281, 334
*ayahuasca* 245–247

Baldwin effect 218
Bayes' rule 39, 80, 189n4. *See also* Bayes' theorem
Bayes' theorem 37–40, 79–80, 101–102, 114, 173, 203, 229, 236n3, 239, 334. *See also* decision theory, design (Bayesian), model (Bayesian), prior, sampling theory
Bayesian prior. See prior
behavior reading 129, 132. *See also* mindreading, theory of mind

bet hedging 161, 165n5
bias 23, 40, 79, 81, 228, 230–234
  conformist 228. *See also* bias (frequency-dependent)
  content 231
  danger learning 234
  error management 81
  frequency-dependent 228. *See also* bias (conformist)
  model-based 230
  prestige 230, 237
  social 79
biological species concept 323
Black Sabbath 187
blindness 23, 112, 125
  change 112
  inattenional 112n4
  instinct 125
  light 23
Broca's area 317
broken wing display 134, 310
brood parasitism 73. See *also* nest parasitism
byproduct 116–119, 142, 163, 330

canalization 56, 63, 156, 171–175. *See also* developmental system (canalized), innateness, landscape (developmental), reaction norm (canalized)
canid 141–144
capuchin monkey 43–49, 102, 150, 224, 227, 245
carbon disulfide 96
celestial navigation 190–191, 227, 256
change blindness 112. *See also* inattentional blindness
chicken and egg problem 220, 326, 329. *See also* first mover problem
chimpanzee 40, 44–46, 66, 130–132, 139–145, 149, 151, 181, 224, 254, 256, 287, 291, 298, 300, 308, 319, 322, 325–327, 331
chimp-human common ancestor (CHCA) 325–326
coevolution 210–223, 248n3, 308, 328, 330, 337
  antagonistic 222
  cooperative 222
  culture-gene 210, 214, 220
  host-pathogen 222
  interspecies 218
  predator-prey 222
  within-species 222, 302, 326
cognition 18–20, 40, 46, 48–49, 57, 61, 69, 85, 87, 129, 146, 152, 211, 215–216, 254–256, 262–272, 279, 283, 286, 289–295, 313, 321, 326, 330, 333, 335–337. *See also* information processing
  causal 255, 289, 331
  emotion versus 87
  function of 254
  general-purpose 80, 265, 277
  goal-directed 195n5
  higher-level 3, 6, 263, 265–271, 276
  lower-level 265–270
  nonlinear 289
  probabilistic 81, 216
  social 106, 126, 141, 255–256, 328
  spatial 93
collaboration 142, 146, 151, 200, 276–277, 280, 286, 292–295, 337. *See also* cooperation, interaction (designed), goal (shared), interface (designed), interface (modular)
  designed 295, 311. *See also* emergence (designed)
communication 45, 54, 59–60, 93, 125, 134, 141–144, 147, 151, 216, 219–221, 239–240, 245, 292, 295–296, 328–331
  conventions of, 312, 318
  cooperative 141–144
complexity 18, 33–26, 45, 55, 94, 102, 112, 144, 170, 204, 219–224, 248n3, 289–291, 298, 302, 314, 318, 327, 335–337
  linguistic 219–223
  ratcheting 224, 243–244, 251, 331
concept 13, 57, 86–88, 102–103, 105, 107–111
  action 120
  artifact 114
  Batman 193
  emotion 152
  food 186
  hunger 93
  mental state 105, 146, 152, 186, 199–202

consciousness 5–11, 23, 40, 47, 48, 128–129, 145n5, 228n1, 238n5, 261–266, 274, 281–287. *See also* phenomenology, qualia, unconsciousness
  bottleneck of 284
constraint 19–20, 23, 31–38, 48, 53, 106, 150, 156–157, 186, 219, 300
  developmental 33, 163
  enabling 31
  evolutionary 32–34, 96, 170–174. *See also* path dependence
  genomic 33
  learning 19, 29–30, 37
  organizational 33
  physical 33
constructivism 49, 315
context dependence 21, 277
contingency 202, 274, 278, 286, 292–296, 319. *See also* detection (contingency), contingent irreversibility, contingent reactivity
  developmental 306–308, 313–315
  spatial 313
  temporal 202, 313
contingent irreversibility 33–34. *See also* constraint (evolutionary), frozen accident, path dependence
contingent reactivity 108, 110, 140
convention 221, 245, 247–254, 268, 296. *See also* coordination device, norm
  neural 296, 312, 314, 318
cooperation 53–54, 134, 141–147, 151, 215, 255, 285, 297, 314, 324, 326, 329–331. *See also* altruism, mutualism
  evolution of 53, 220
  norms of 328
co-optation 307. *See also* byproduct, leveraging
coordinates 29–30
coordination 126, 137, 143, 219–221, 247, 253, 271, 280, 287, 302, 313, 328–330. *See also* convention, coordination device, games (coordination), interaction (coordinated), problem (coordination)
coordination device 221
cortex 3–4, 30, 302–307, 315–317, 328–329, 333
  auditory 2, 306
  dorsal medial prefrontal 146
  frontal 146
  inferotemporal 294
  medial frontal (MFC) 146
  neo- 3n3, 139, 321, 328
  occipitotemporal 294
  orbitofrontal 295
  prefrontal 294
  somatosensory 316–317. *See also* somatosensory homonculus
  somatomotor 316–317
  visual 65n4, 137, 187, 294, 317
cortisol 314
corvid 139–145
counter-gradient selection 173. *See also* antagonistic selection
cue 20, 57, 61–62, 70–73, 80, 88, 92, 101–103, 108–115, 120, 129, 132, 136, 140, 227–236, 240, 248–249, 255, 280, 290, 311. *See also* index, information, signal, validity (cue)
  animacy 122
  communicative 143–145
  contextual 279
  motion 108–110, 129, 278, 309
  pedagogical 240–241
  perceptual 102, 135, 195, 199, 293
  prepared 234
  proximal 96
  signal versus 59
  social 233, 240, 309
  static 112–113
  visual 92, 100

danger 6, 104, 229n2, 231–235, 284–285, 294
Darwin's syllogism 159–164, 228
decision making 3, 6–7, 18, 48, 79, 89–94, 101, 105–106, 120, 184, 204, 265, 269, 275, 280, 285–286, 293–294, 309
decision theory 81, 224n3
demon 271. *See also* module, pandemonium
descent with modification 1, 8, 32, 34, 90, 287, 291, 299–300, 308, 325, 332, 336. *See also* homology

design 5, 18–19, 27–28, 34, 52–54, 62–66, 69, 77–79, 90, 94, 116, 134, 136, 140, 158–176, 185–189, 194–198, 202, 209, 212, 215, 223, 270–273, 285–287, 291–299, 305, 316, 337
algorithmic 273
ancestral 300–301
assembly-line 269
Bayesian 173
brain 147, 219, 325, 328, 336
canalized 171
complex 35, 309
computational 133
culturally evolved 245–252
developmental 10, 32, 55, 126, 158, 160, 165, 178, 193, 290, 307, 318
distributed 35
duplication and divergence in 301–302
evolved 10, 151–152, 262, 318, 332
feature 10, 24–25, 29–30, 33–34, 74, 97, 100–101, 107, 113, 118–119, 150, 161, 169, 180, 190, 204, 225–228, 236–241, 263–267, 273–277, 279, 282, 285, 302, 303
hierarchical 300, 306
hub-and-spoke 295
information-processing 302
inheritance of 311
interaction by 317
organismic 22–23, 52, 55, 155, 160, 291, 297
phenotypic 176, 203, 307, 325
reincarnation 159, 159n2, 161–163
replication 162
space 59, 76, 100, 129, 161–172, 219, 297
detection 18, 20, 61, 64–68, 70, 102, 107–119, 227, 240, 270, 278, 293, 309–312. *See also* recognition
agency 126, 146, 200, 309
animacy 107–114, 134, 278
antigen 218
change 113
commonality 310
contingency 201–202
edge 67, 275, 316
emotion 278
eye 134, 310
eye direction. *See* detection (gaze)
face 107, 115–118, 278, 295–296, 310–312
feature 272
food 95, 101, 195. *See also* detection (nutrient)
gaze 134–137, 199. *See* detection (eye direction)
human 115
identity 117–118
light 21–26. *See also* phototaxis
mother 73
motion 108–112, 134, 309
nutrient 195–197, 198n6. *See also* detection (food)
object 88, 304, 310
odor 268
pattern 89
predator 309
signal 81, 108
snake 6, 7, 9
detector. *See* detection
development 8–11, 25, 28–29, 32–35, 39, 50–56, 60–62, 71–72, 75–76, 91, 100, 114, 118n6, 125–126, 136, 144, 150, 152, 155–208, 226, 262, 275, 287, 289–291, 296–298, 303–319, 324, 333–338
ballistic 19
body 29–30, 157, 298, 314
brain 11, 19, 29–30, 188–189, 202–204, 303–306, 313, 333
canalized. *See* canalization
closed 179
cognitive 19
contingent 30, 305–306, 308
eye 64n3
guided. *See* development (targeted)
hierarchical 303–306, 319
open 179, 182–208
phenotypic 32, 173
reliable 126, 191, 202
targeted 19–20, 192, 308
developmental conversion 179, 185
developmental system. *See* development, system (developmental)
differentiation 287, 301–308, 315, 329. *See also* module spawning, parcellation

content-based 306. *See also* differentiation (what-based)
niche 301. *See also* adaptive radiation
ontogenetic 301
what-based 305–306. *See also* differentiation (content-based)
when-based 305–306
where-based 305
dimorphism 325, 327–328
disequilibrium 77–78, 204. *See* adaptive lag
dissociation 118, 287
divergence 287, 299–301, 307, 311, 319. *See also* homology, orthology
diversity 4, 8, 30, 120, 170, 199, 209, 248n3, 287, 320, 324
adaptive 267
mental 139, 336
division of labor 4, 74n6, 115, 120, 196, 200, 263, 270, 275, 278–279, 293–296, 330, 337
dog. *See* canid
domain 5–7, 26–28, 95, 101, 106, 107n1, 141, 186, 215, 253–257, 301, 306
actual 27, 118–119, 180, 249
generality 5–7, 19, 80, 229, 257, 315, 337
learning 95, 102, 235
narrowness 5–7, 265–267
proper 27, 118–119, 196–197, 211, 214, 217, 222, 225
source 195n5
specificity 3, 26–27, 30, 76, 115–119, 121, 148, 191, 199, 214, 264, 273–274, 279, 303, 306, 308, 314–318, 336
dopamine 195n5, 314
dorsal stream 74
drift 50, 52, 62, 220, 247, 301, 325
dualism 7, 257, 265–266, 281, 286. *See also* model (dual systems)
duplication 150, 298–300, 310–311, 325–326, 329. *See also* homology (serial), paralogy
and divergence 299–300, 307, 311
gene 300–301
dynamical system 212, 222, 242, 256–257, 275, 290, 326, 337. *See also* dynamics
dynamics 29–30, 159, 211–214, 235, 277, 284, 287, 289, 291, 295, 313
cultural 235
developmental 290
evolutionary 164, 211–214, 226, 242, 255–256
feedback 230, 311
frequency-dependent 213
predator-prey 214
replicator 212
runaway 222–223, 230
self-stabilizing 222–223
thermo- 59

ecology of mind 291
EEA. *See* environment of evolutionary adaptedness
egg timer neurons 136–138, 150
emergence 30, 271–272, 284, 286, 289–290, 294, 302, 337
designed. *See* emergence (selected)
developmental 30, 157–159, 303–306
selected 272, 277–281, 290, 306, 308, 311, 315, 318
emic / etic distinction 238n5
emotion 2, 9, 87, 90–91, 115, 129, 152, 242, 252–253, 261, 263, 265n2, 275, 295, 309, 311, 331–332. *See also* detection (emotion), motivation
anger 255, 275, 278
cognition versus 87
compassion 331
disgust 92–93, 97, 100
embarrassment 145, 331
expression 125
fear 2, 6, 9, 91, 109, 183, 232, 233–234, 275, 277–278, 294, 309
happiness 277–278
hunger 2, 60, 91–93, 100, 136, 179, 284–285
moral 253
social 331
worry 278
empiricism 28, 156
emulation 42. *See also* imitation, goal imitation
encapsulation 264–283
entropy 59

environment 5, 10, 18, 20–25, 32, 45, 50–57, 60, 66, 72–73, 77, 79, 93–97, 112, 150, 157, 166–184, 188–189, 200–201, 209–215, 241–243, 247, 290, 297, 313, 327–328
adaptively relevant (ARE) 170
ancestral 6, 77, 119, 194, 196, 233–4, 266
cultural 227–232
developmental 56, 126, 144, 155, 165, 174, 178, 183, 235
external 29, 159, 181, 313, 315
of evolutionary adaptedness (EEA) 25, 77, 167–175, 180, 188, 191, 197, 225, 229
internal 29, 159, 181, 313, 315
normal 25
novel 11, 166–169, 194, 198, 203–204
probabilistic 79, 81, 161, 210, 262
social 177, 232, 311
space. *See* space (environment)
species-typical 25
epigenetics 32, 56, 170, 172, 297, 310, 325, 328
episodic buffer 286. *See also* memory (episodic)
equipotentiality 91
cortical 315
cultural 235
error 59, 109, 160, 162, 168
management 81, 108, 236. *See also* bias
sampling 231
threshold 160, 246
trial and 35, 43–44, 49, 126, 169, 246
Type I 81, 137. *See also* false positive
Type II 81, 137. *See also* false negative
essentialism 321–323
evo-devo. *See* evolutionary developmental biology
evolution 1–84
autocatalytic 29, 53, 222, 230, 327–328
convergent 64, 140n4, 145n5, 146, 210
cultural 217–252, 308, 331, 337
cumulative 18, 35, 49, 53, 63, 68, 76, 220, 224–226, 243–247, 297, 331, 337 *See also* additivism, evolution (ratcheting), hill-climbing
ratcheting 34–35, 50–53, 224, 243, 246, 248n3, 251, 326–327, 331, 337 *See also* evolution (cumulative), hill-climbing
Evolutionarily Stable Strategy (ESS) 213
evolutionary developmental biology (evo-devo) 32, 157, 287, 300–301, 307, 315
evolutionary psychology 1–2, 11–12, 28, 65, 77, 89, 119, 261, 331
alternative hypotheses in 119, 167
caricatures of 77, 204, 264
criticism of 1–2, 28, 77, 89, 256, 287
holistic 11–12, 262
just-so stories in 2, 334
experience projection 140
explanatory walling-off 261
extrapolation 188–189. *See also* interpolation
eye 61–68, 112, 224, 245, 271, 284, 293
compound 172
cue 110, 116–117, 129, 134–135, 278, 295, 310
cup 64, 68, 76
detection. *See* detection (eye), detection (gaze)
image-resolving 64–65, 76
lens 64–65, 68, 172
spot 22n2, 61–63, 134

facial fusiform area (FFA) 303, 310, 317
false belief tracking. *See* tracking (belief)
false negative. *See* error (Type II)
false positive. *See* error (Type I)
felicity condition 24, 119, 197–198, 204, 255, 337–338. *See also* novelty (felicitous)
first law of adaptationism 8–12, 23, 25, 27, 39, 79, 100, 158, 167, 170, 176, 180, 211, 267, 269, 302, 336
first mover problem 220. *See also* chicken/egg problem
fit 8, 18, 25, 27–28, 45, 57, 87, 91, 97, 105, 144, 186, 190, 193, 198, 217, 225–226, 247, 226
form-function 8, 23, 71, 74, 158, 266–269, 273
goodness of 170

mind-world 7–9, 10–11, 24–25, 28, 49, 57, 68–69, 90, 97, 100, 106, 167, 169–170, 176, 179, 229, 308, 337
fitness 6, 9, 48, 51–56, 59–70, 90, 97, 99, 104, 108, 133, 159–160, 212–213, 221–223, 244, 252–254, 256, 263, 268–269, 275, 283 -284
badness 25, 31, 49, 54, 80, 184, 228, 253
benefit 18, 21, 49, 74–75, 77–79, 137, 142–143, 147, 202, 219, 230–231, 242, 247, 275, 330
cost 73–75, 77, 81, 142–143, 275
function 219
goodness 25, 31, 34–35, 49, 54, 61, 97, 112, 134, 164, 169–170, 173, 184, 203, 211, 228, 230–233, 236, 268, 285, 310–311
hill 51–54, 59–68, 99, 134, 223, 244, 291
increase 18, 22, 25, 64, 77, 86, 96, 100, 165, 191, 200, 236, 297, 310–311, 328
landscape. *See* landscape (fitness)
mistake 78
optimum 52, 172–173
relevance 100, 114, 147, 160–163, 268, 284, 309
rightness 23, 254
space. *See* space (fitness)
usefulness 95, 227–229, 234
variation in 32
flexibility 5–6, 11–12, 18–23, 31, 35, 76, 81, 89, 94, 106–107, 155, 127, 130, 141, 145, 166, 173, 299, 204, 215, 261–280, 286, 196, 307, 318, 330. *See also* plasticity
food 17–18, 43–47, 68, 75, 79, 88, 90–101, 131–132, 143, 146, 150, 178, 179, 183, 194–198, 209, 225, 235, 239n6, 297, 313, 320, 328
aversion 91, 101
caching 73, 139–141
competition 139–144, 237
concepts 90–101, 186
craving 93, 204
evaluation 98–99, 195–196
learning 91–101, 211
ontology of 90
preferences 77, 97, 224, 232, 284
taboo 97
foraging 43, 93–94, 96–97, 102, 178, 195–197, 313–314, 328
theory 93–94, 195n5
fovea 112
frame shift 98
free will 261, 281
frozen accident 33. *See also* constraint (evolutionary), contingent irreversibility, path dependence
function 3–4, 8, 10, 17, 47, 57, 64, 67, 88, 91, 106, 121, 125n8, 134, 138, 159, 193, 195–196, 219–220, 236, 244, 254–255, 267, 273, 275n4, 282–285, 301, 304–305, 308, 313–317, 330. *See also* fit (form-function)
artifact 114, 247–251
conventional 248–251
decay 136–138
input-output 183–186, 189, 264
likelihood 80
logic 187
mapping 55–56, 62, 139n3, 160, 163, 166, 170, 172, 175, 177, 179, 182–183, 185–187, 191, 203
mathematical 67, 183, 212, 228, 289
time discounting 273
functional fixedness 250. *See also* problem (candle)
functional incompatibility 74
functional morphology 56–57

game 211–215, 248, 322
against nature 214–215, 232, 242
against people 214–215
coordination 220
host-pathogen 214
economic 151
language 220
predator-prey 214
theory 147, 211–215, 220, 226, 252, 255
gene 157, 160, 162n4, 172, 210, 214, 217, 219–220, 222, 310, 314, 322, 325
cyclic AMP-dependent protein 314
cyclic GMP-protein kinase 314
duplication 299–301

gene (*Cont.*)
frequencies 50
*Hox* 314
*Pax-6* 64n3
polymorphism 322
space. *See* space (gene)
toolkit 301
gene expression 29, 32, 177–178, 285n5, 287, 298–299, 305, 314–315. *See also* epigenetics, gene regulation
changes in 177, 287, 298–299
contingent 29–30
differential 305, 308
experience-dependent 178, 313
species differences in 319
gene regulation 29, 179n10, 318. *See also* epigenetics, gene expression
changes in 150, 217, 219
contingent 29, 184n2, 313–314
differential 29, 301
species differences in 325
gene shortage argument 157
general-purpose mechanism. *See* mechanism (domain-general)
genetic drift. *See* drift
gestalt 67, 108
goal 17–18, 99, 102, 120–125, 128–129, 237–241, 245–249, 277, 295, 330–331
directedness 46, 89, 108–111
imitation 42–49, 50, 60, 89
opacity 46, 238
shared 146
state 46, 238–239, 295
strategic 222
grammar 36–38, 47, 215–221, 307, 324, 331
acquisition 324
universal (UG) 37–38, 215–216. *See also* Language Acquisition Device
greeble 118–119

hard-wiredness 274–276
heritability 162
hill-climbing 18, 32, 35–36, 50–81, 99–100, 135, 162–163, 185, 200, 211, 219, 223, 243–247, 270, 291, 302, 310
homolog. *See* homology
homology 46, 64, 90, 97, 138–139, 145, 298–300, 311, 317, 325–326, 330, 332
serial 298–300. *See also* paralogy
hopeful monster 311
host 73, 142, 198, 209, 212–214, 222. *See also* parasite, coevolution (host-pathogen)
human nature 256, 263, 320–332
human uniqueness 46, 319–335
hunger. *See* emotion (hunger)
hunting 99, 142–144, 327–329

imitation 41–49, 124–125, 239, 295
blind 42. *See also* mimicry
goal 42–49, 50, 60, 89, 124, 238. *See also* emulation
over- 45, 236, 238
immune system 3, 210, 218, 245
immutability. *See* assumption (immutability)
imprinting 73
indeterminacy 38, 40
of translation 38. *See also* problem (gavagai)
referential 38. *See also* problem (gavagai)
index 88, 129, 132, 134, 136–143, 199, 231–232. *See also* cue, signal
indigo bunting 190–194, 226, 256. *See also* celestial navigation
inductive bet 23–28, 35, 38–40, 48, 56, 77, 79–81, 89–90, 95, 99, 104, 114, 119, 121, 160, 169, 183, 187, 189–194, 197–204, 211, 215, 218, 221–222, 227–229, 233, 236, 238–240, 242, 247, 250–251, 254–255, 274–275, 280, 293, 296, 307, 309–312, 322, 337. *See also* felicity conditions
infant-directed speech 240. *See also* motherese
inference 24, 25, 31, 41–47, 59–61, 64, 66–67, 71, 86, 89–90, 98–100, 105–107, 119–120, 128–130, 134, 137, 186, 237, 240–241, 249–251, 268–271, 277–280, 294–295, 320
causal 246, 331
deductive 24, 39
general-purpose 37

inductive. *See* inductive bet
intentional 107–109, 128–145, 199. *See also* mindreading, theory of mind
perceptual 110
probabilistic 107, 231
rule 106, 122, 188
system. *See* system (inference)
inferential opacity. *See* opacity
information 1–11, 18–20, 23, 27–30, 41, 45, 48, 61, 63–72, 74, 85–104, 105–120, 135, 150, 157–160, 172, 180, 186, 194–199, 217, 243, 247–248, 264–268, 273–280, 292–296, 302–306, 309–311, 314, 317–318, 331, 336
attractors 235
conscious 282–284
cultural 225–236, 250
environmental 192
neural 187, 268–273
processing. *See* processing (information)
Shannon theory of 26, 57–60, 80, 129, 135–139
shapes 57
signature 279, 306. *See also* tag (informational)
space. *See* space (information)
spatial 76
structures 30, 40, 114
tag. *See* tag (informational)
transmission 61, 63, 143, 210–211, 224–225, 231–232, 236, 239–240, 246
innateness 5–10, 19, 26, 28–29, 32, 56, 63, 81, 89, 114–115, 117, 126–127, 151, 155–161, 166, 174–175, 180, 183n1, 191–193, 204, 210, 215, 218, 223, 262–266, 271, 281, 286, 290, 303, 307, 335–337. *See also* canalization, reaction norm (flat)
appearance of 192
inparalog 300, 309, 326. *See also* homology, paralogy
input 5, 26–29, 36–39, 61–62, 66–67, 69, 88, 95, 114, 121, 183–196, 211, 216, 225–226, 264, 267, 272–274, 278–279, 293, 302–313, 318
analyzers 102–103
condition 109, 226, 229, 231, 235, 272–274, 278, 294–297
criterion 103, 197, 274, 293
cultural 229, 318, 337
environmental 10, 157
fuzzy 272–274
novel 194–196
perceptual 105–107, 335
system. *See* system (input)
instinct blindness. *See* blindness (instinct)
interaction 11–12, 289–332
between systems 3, 6, 45–46, 88, 101,106–107, 115, 120, 126, 129, 137, 145n5, 147, 152, 204, 212, 242, 264, 267, 270–271, 279–280, 286, 289–291, 293–298, 302, 310, 326, 330, 337–338
coordinated 330
contingent 278
designed 66, 263, 277, 318, 333
developmental 157–158, 176, 222, 298
genetic 158–160, 168, 176, 241, 305, 306, 315
human-artifact 245–252
modular 272–273, 279–281, 285, 289–291, 317
social 116, 130, 137, 150, 215, 243, 254–256, 280
third-party 253–254
with environment 25–26, 29–31, 40, 56–57, 65, 75, 79, 85–86, 99, 126, 144, 149, 157, 168–169, 181, 183, 193, 197–199, 201, 232, 243, 256, 262, 275, 291, 302, 304, 315, 328
interface 87, 101, 151, 331
designed 151, 337
problem 273
interpolation 188–189. *See also* extrapolation
intersubjectivity 41, 128–129, 146. *See also* mindreading, theory of mind
inversion effect 117–118. *See also* processing (configural)
islands of competence 122, 145

kin selection 104, 220

lac operon 178, 313
landscape 49–55, 80, 94, 195, 201, 209, 235, 256
  adaptive 50–55, 60, 64n3, 76, 150, 162–163, 200–202, 214, 256, 291, 330
  developmental 55
  evolutionary 49
  fitness 52, 59, 63–65, 167–173, 176, 212, 312. *See also* landscape (adaptive)
  phenotypic 53
  rugged 52, 173
language 3, 7, 9,13, 19, 31, 36–40, 48, 53, 59, 60, 65, 66, 80, 90, 97, 103, 129, 138n2, 145n5, 147, 165–166, 171, 188–189, 197, 212, 214–227, 236, 242, 245, 247, 251, 254, 257, 283, 292, 307, 317, 321–322, 324–332, 335, 337
  acquisition. *See* learning (language)
  acquisition device (LAD) 37–38, 215–218
  learning. *See* learning (language)
  sign 125
learning 5–11, 19, 28–30, 45, 48, 66, 79–80, 90–91, 101, 106, 117, 127, 129, 137, 140, 156–157, 172, 180, 184, 189n4, 191, 204, 227–231, 234, 308, 328, 331
  artifact 114–115, 248–252
  associative 92–93, 101
  conformist 229
  cultural 232–247
  danger 233–234
  domain-general 19, 36, 41, 91, 155, 229, 313, 332, 335–337
  domain-specific 27, 91, 94–95, 99, 102, 114, 196, 308
  experience-expectant 95, 190, 202. *See also* learning (prepared)
  fear 91, 232–233
  food 91–97, 99, 101, 211, 232
  general-purpose. *See* learning (domain-general)
  gene expression and 178
  guided 95, 190
  individual 94, 96, 232–233, 247
  language 36–39, 98, 147, 215–221, 226
  prepared 95, 190, 231, 234
  social 43–45, 94–97, 124, 147, 151, 232–233, 329–330. *See also* cultural transmission
  specialized. *See* learning (domain-specific)
  statistical 41, 216, 335
  tool. *See* learning (artifact)
  trial-and-error 44
leveraging 25, 34–35, 57, 95–96, 104, 107, 109–110, 138, 237, 245, 300. *See also* co-optation
light box 124, 240–241, 248
limb 4, 30, 57, 157, 174, 287, 298–301, 304, 314–315, 318
logistic growth 183, 213
looking time paradigm 131–132, 135. *See also* violation of expectation paradigm
lookup table 185–189, 203

Machiavellian intelligence hypothesis 147, 222–223, 255
maladaptiveness 78–81, 197, 204, 230–231
map 304
  macro- 304
  meso- 304
  micro- 304
mapping function. *See* function, mapping
Margaret Thatcher illusion 117. *See also* inversion effect
marginal value theorem 94
mechanism 2–3, 5–7, 28–29, 35, 37, 42, 45, 53–54, 63, 72–76, 81, 85, 92, 95, 106–107, 111–119, 122, 126–128, 145n5, 169, 182, 193, 202, 211, 225–241, 255–256, 268–273, 283–285, 328–331, 334–338
  brain. *See* mechanism (mental)
  causal 33, 184, 245
  cognitive. *See* mechanism (mental)
  cultural transmission 225–246
  developmental 25, 35
  domain-general 5–7, 19, 26–27, 32, 80, 101–102, 204, 267
  domain-specific 5–7, 26–27, 46–48, 72–76, 101–102, 111–119, 128–129, 132–151, 171, 195–199, 204, 286, 292–293, 326

general-purpose. *See* domain-general
higher-level 271
information-processing 27, 60, 69, 80, 273
learning 10, 29, 41, 95–97, 101, 106, 114, 190, 216–223, 225–241, 246–253
lower-level 5, 271
mental 11, 32, 40–41, 57, 60, 87–91, 106, 111–119, 132–151, 156, 195–199, 214, 246, 254, 261, 263–264, 274–281, 286–287, 291, 294–297, 300, 303, 306, 310–312, 317, 334
perceptual 2, 21, 23, 25, 27, 266, 268, 271
psychological. *See* mechanism (mental)
specialized. *See* mechanism (domain-specific)
meme 235
memory 30, 73–75, 86–87, 92, 94,107, 135, 139, 195n5, 202, 235, 248, 273, 294, 314
episodic 74
long-term 234
semantic 74
working 7n5
mental adaptation. *See* adaptation (mental)
mental architecture 109, 242, 262, 267, 269, 273, 287, 311, 314, 329, 337
mental mechanism. *See* mechanism (mental)
mental state 41, 47, 89, 105, 116, 128–130, 132, 134–135, 137–138, 141, 144–152, 172, 183,186, 199, 202–203, 278, 317, 331. *See also* mindreading, theory of mind
mental time travel 74
metaphor 25, 36, 50–51, 55, 57, 159, 168, 173, 189, 211, 220, 222, 256, 265n2, 270n3, 272, 279–280, 284, 292, 294, 295, 315, 331
metarepresentation 147–148. *See also* M-representation
microRNA 305, 315
mimicry 42, 134. *See also* imitation
mindreading 41, 46n7, 70n5, 128–140, 144–152, 225, 230, 262, 292, 324–326, 330–331. *See also* theory of mind
in corvids 141
in dogs/canids 141–144
mirror neuron 124–125, 237, 317
miss. *See* Type II error
mode of construal 89, 98–99, 321. *See also* intuitive ontology
model 6–7, 46, 57, 93, 96, 114, 119, 135, 156–157, 159, 184n2, 239, 269, 285, 292, 301, 328, 331–332, 334–337
ballistic 19
Bayesian 38–40, 79–80, 101–102, 236n3, 239, 334
bucket brigade 107
bulletin board 279, 293–295
causal 331
computational 273
cultural evolution 225–233
direct perception 46n7
dual systems 265, 281, 286, 335
dynamic programming 189
dynamical systems 275n4
emergence 272, 281
enzyme 294–295
error threshold 160
evolutionary 217–220
folk 281, 287, 312
game theory 213
hierarchical 107n1, 306–308. *See also* pandemonium
hill-climbing 246
igloo 265, 336. *See also* systems (dual), systems (higher / lower), systems (central / peripheral)
Lego 264, 276, 286–287, 313–315, 318
lens 57
linear 289
mirror 57
neural democracy 285–286. *See also* pandemonium
neural network 275n4, 279, 285, 304
null 34n5
optimality 94
pandemonium 272–274
pipeline 270, 279
probabilistic 79, 216
reaction norm 187, 203
scissor 57
subfunctionalization 301
workspace 279, 286, 295

modularity 3, 174n9, 264–267, 281, 286, 298, 307, 315–318, 330
  massive 3, 298
  network 329
modularization 303–307, 316, 329. *See also* module spawning, parcellation
module 3–5, 89, 257, 264–272, 277, 280–281, 286–287, 289–290, 294, 298–303, 317–319, 326, 329, 336
  Fodorian 7–9, 264–271. *See also* model (igloo)
  spawning 114, 119, 304–311, 319. *See also* modularization, parcellation
modus ponens 187–188
morality 147, 151, 252–256, 280, 324, 328. *See also* norm (moral)
morph 176–179
  contextual 179, 287
  facultative 176–178. *See also* polyphenism
  fixed 176–177. *See also* polymorphism
morphogenetic field 29
morphospace 51, 56
mosaic 8, 145n5, 297, 336
motherese 240. *See* infant-directed speech
motivation 87–92, 104, 147, 240, 248–249, 252–253, 275, 331. *See also* emotion
motor control 2, 6, 7, 263, 265, 276
moving target 209–223, 225
M-representation 148–151, 200. *See also* metarepresentation
mutation 18, 22, 32–33, 50, 52, 62, 97, 150, 161n3, 162, 200, 202, 218, 220, 297–299, 310–311, 322, 325, 329
mutualism 141–142. *See also* cooperation
  byproduct 142

nativism 19, 28, 156, 223
  starting-state 19
natural kind 321
natural selection. *See* selection, natural
nature / nurture dichotomy 29, 126, 156, 241
nausea 91, 95, 101
nest parasitism 73, 76–77, 119, 198. *See also* brood parasitism
neural democracy. *See* model (neural democracy)
neural reuse hypothesis 307
niche construction 54
niche differentiation 301. *See also* adaptive radiation
noise. *See* error
norm 105, 215, 232, 252–254, 256, 307, 328, 337
  conventional 252–253
  moral 252–253
  reaction. *See* reaction norm
  social 221, 245, 252, 337
novelty 10–11, 81, 189, 193–194, 204, 225, 299–300, 326, 337
  aspectual 194
  felicitous 198, 307
  paradox of 25, 262

object 17, 30, 38–40, 48–49, 65, 80, 85–91, 93, 100–101, 105, 115–124, 131–133, 138, 145–146, 171–172, 187–188, 194, 196–198, 227, 233, 235, 239–240, 247–252, 268–271, 274–275, 278–279, 291
  categorization 70, 85–86, 98–99, 110–111, 134, 196, 309
  file 67, 108, 293, 308
  mechanics 109, 211
  parsing 66–68, 75–76, 88, 101, 107–108, 109n2, 111, 114, 126, 271, 293, 309–312
  permanence 69–73, 76–77
  recognition 111–113, 115–119, 293–296, 304, 306–312, 316–318. *See also* face recognition
  template. *See* template (object)
  tracking. *See* tracking (object)
Occam's razor 334. *See* parsimony
one reason theories of human uniqueness 319
ontological commitment 87–89, 105–106, 109–110, 117, 121–122, 127, 225, 250. *See also* assumption
ontological correctness 322
ontology 57, 87–90, 124–127, 225, 250, 321–322
  artifact 250
  food 90–98
  intuitive 87–89

linguistic 90
mental state 89
object 88
of groups 90
social 90, 104–127
substance 98–99
tool 90
opacity. *See* problem (opacity)
open content 184, 187, 188n3, 191, 194, 210
operation 5, 102, 187, 211, 270–271, 273, 281, 312–313, 318
carry-through 187
context-specific 276
domain-general 5
logical 102
opsin 21–27, 34, 40, 59–64, 299, 301–303
organization 3, 115, 126, 248, 262, 287, 291, 336
complex 129, 287, 295
constraints of. *See* constraint (organizational)
cortical 3, 290, 315–316, 319
domain-specific 315–316
functional 7, 152, 316
hierarchical 107n1, 287–288, 300, 304–305, 315–316
higher-level 106, 272
modular 298
self- 271, 289, 305, 337social 326
ortholog 300. *See also* homolog
ostension 240. *See also* pointing
outparalog 300. *See also* paralog
output 61–62, 69, 74, 85, 88, 106, 148, 183–189, 203, 226, 237, 264, 269, 271–275, 279–280, 283–285, 293–295, 297, 309, 312–313
overimitation. *See* imitation (over-)
oxytocin 314

pandemonium. *See* model (pandemonium)
paradox of novelty 25, 262
paralog 298, 300–301, 307, 309–310, 325–326, 329. *See also* duplication, homology, inparalog, outparalog
paralogy. *See* paralog
parasite 73, 76–77, 119, 198, 210, 213, 214n3,. *See also* coevolution (host-pathogen), host, pathogen
parcellation 304, 329. *See* also modularization, module spawning
parental investment 314, 326–327
parsimony 8, 100, 133, 332, 335–336. *See also* Occam's razor
parsing 66
action 120–127, 237–238
animate-inanimate 108, 112
artifact 248
body movement 125
conceptual 86, 106
facial expression 125
goal 121, 331
language 283
object 67–88, 109n2, 111, 271, 293, 309–312
speech 264–265
path dependence 33–34, 51–53, 63, 76, 200, 291–292, 296–297, 301, 305, 308, 311–312. *See also* evolutionary constraint, frozen accident
pathogen 92–93, 209–218, 222, 245. *See also* coevolution (host-pathogen), host, parasite
pedagogy 237–241, 249, 251
perception 2, 6–11, 27, 41,48, 67, 70–75, 77, 85–87, 90, 93, 102–103, 105, 123–126, 134–135, 140, 184, 195, 199, 242, 261, 263, 266, 268–276, 287, 293, 295, 317, 329, 335, 337
direct 46n7
inferential nature of 41
motion 107–115
speech 121
peripheral system. *See* system (lower-level)
peripheral vision 113–114
personality 177, 179
phenomenology 48, 93, 282. *See also* qualia
phenotype 10–11, 28, 32, 35, 50–56, 62, 87, 126, 157–184, 192, 199, 203–204, 211, 218–219, 228, 243, 262, 290–291, 297–303, 307, 312–315, 318–319, 324–325, 336–338. *See also* trait

phenotype (*Cont.*)
canalized. *See* canalization
cultural 228
landscape. *See* landscape (phenotypic)
mental 11, 90, 304
plastic. *See* plasticity
space. *See* space (phenotype)
phonological loop 286
phototaxis 20–23, 34, 61, 80, 110, 214, 355
pipeline system. *See* model (pipeline)
plasticity 3, 30, 56, 62, 156, 172–176, 202–204, 209, 262, 302, 317, 321, 328. *See also* flexibility
adaptive 161–166
developmental 176–179
psychological 179
pointing 58, 143, 240. *See also* ostension
polymorphism 176, 322. *See also* morph (fixed)
polyphenism 177. *See also* morph (facultative)
population genetics 55, 226
pop-out effect 109, 113, 227
poverty of the stimulus 36. *See also* problem (opacity)
predator 17–18, 52, 68, 70, 77–78, 90–91, 108, 134–136, 144, 171, 209, 212, 214, 222, 224, 233–235, 309–310
preformationism 155, 158, 204, 307, 315
pregnancy sickness 92–93
prespecification 156
priming 279–280
prior 37–39, 79–80, 101–102, 114, 236n3, 334. *See also* Bayes' theorem
problem 1–16
adaptive 65–67, 70–73, 91, 100, 115, 117, 120–121, 141, 172, 231, 239, 246, 266, 268, 273, 336
candle 250
coordination 219–221, 247
frame 36, 40–44, 115, 120–122, 174, 226, 228, 236, 239–241, 246, 250. *See also* problem (opacity)
gavagai 38–40, 47, 101, 147, 239–240, 331
interface 273
opacity 36, 147, 215–216, 230, 236, 238–239, 245. *See also* problem (frame), problem (gavagai), poverty of the stimulus
processing 273, 293
the 1–16
processing 18, 85, 107, 116, 156, 195, 224, 226, 269, 287, 317
action 120, 125. *See also* parsing (action)
byproduct 119
cascade. *See* processing chain
chain 120, 270, 277–280, 283, 294–296, 318
collaborative 295
conceptual 102, 107
configural 117. *See also* inversion effect
emotion 311
face 117–118, 310–311, 317
food 195
hierarchical 107n1, 110, 275
holistic 117–118
information 18, 20, 27, 39, 50, 60, 63, 68–69, 76, 80, 86, 103, 120, 210, 229, 267, 273, 292–294, 302
logic 235
motion 293
neural 289, 312
object 291
parallel 113, 278–281, 284, 294–295, 337
perceptual 269
problem. *See* problem (processing)
serial 113, 295
strength of 274
unconscious 283
visual 66–67, 74, 111–112, 278, 316
propositional attitude 148
prosopagnosia 118
psychological dualism. *See* dualism

qualia 285. *See also* phenomenology
quorum sensing 285n5

random walk 23, 35, 50
biased 23, 40
guided 35

rare-type advantage. *See* selection (frequency-dependent)
ratcheting 34–35, 50, 53, 224, 243, 246, 248n3, 251, 326–327, 331, 337. *See also* evolution (cumulative), evolution (ratcheting), hill-climbing
rationality 25, 121, 204
  assumption. *See* assumption (rationality)
  ecological 25
rat 91–96, 232
reaction norm 55–56, 95, 163–167, 170, 181, 183–193, 262, 275, 285, 296, 308, 316–317, 319, 324, 328–329, 337
  adaptive 165, 167
  canalized 56. *See also* reaction norm (flat)
  classical 183–184, 203–204
  closed 203–204
  flat 56, 62, 170–175. *See also* reaction norm (canalized)
  non-adaptive 165
  open 184–193, 203, 211, 215, 303–306, 337
  plastic 56, 165, 170, 175–179, 328
reasoning 3–7, 27, 48, 86–87, 120, 144, 249, 261–265, 318, 331
recognition 66, 77, 111–119, 211, 272. *See also* detection
  emotion 280
  face 116–119, 127, 151, 279–280, 291, 293, 303, 310–311, 317
  ignorance 133
  kin 211
  letter 272
  object 111–112, 119, 293–294, 304, 307–311, 316, 318
  offspring 77
  pattern 66, 272
Red Queen 210. *See also* arms race
reliable development. *See* development (reliable)
replication fidelity 160–162, 166, 246
replicator 160–162, 212, 242
replicator dynamic 212
representation 17, 23, 31, 41, 45–49, 57, 60, 63, 65–71, 79–80, 85–89, 93–96, 100–103, 105–107, 111, 115, 138, 150–152, 184, 188, 195–196, 199–202, 238n5, 248n3, 256, 268, 274, 276, 279, 282–284, 293–296, 303, 306, 331, 337
  belief 140–141, 145, 148–149
  composite 150, 200–201, 279, 318
  cued 70n5
  detached 70n5, 136
  format 48, 60, 103, 107–108, 141, 148–149
  iconic 137
  indexical 138, 141
  innate 89, 156–157
  M- 148–151, 200
  mental state 47, 107, 129–130, 133, 135, 137, 140–141, 146, 200–201, 331
  meta- 147–149
  object 67, 85, 108, 112, 271, 293, 306–307, 309, 312. *See also* object file
  recursive 149
  type 48, 60, 195
reputation 78–79
restraining bolt 37–38
retina 5, 64–69, 85–88, 112, 133, 187, 190, 270, 274
rewiring 149–150
robustness 109n2, 198, 280, 307
rule 37, 44, 57, 80, 100, 109, 197, 283
  Bayes' 39, 79–80, 189n4. *See also* Bayes' theorem
  computational 71, 137
  cultural 307
  decision 70, 141, 195
  fuzzy 273
  grammatical 216–218, 221, 226, 236, 245, 331
  heuristic 239. *See also* rule of thumb
  if-then 23, 102, 184n2, 187
  inference 39, 105–106, 122, 188
  genetic 3n3
  information-processing 39
  learning 29, 36, 204, 337
  logical 188
  marriage 252
  moral 324
  of thumb 130, 132, 135. *See also* rule (heuristic)
  optimal foraging 102

rule (*Cont.*)
parsing 67
processing 293
stopping 37–38
transmission 226
wiring 302, 305–306

Sally/Ann task 131–132
sample size 229, 231
sampling theory 229, 232. *See also* Bayes' theorem
schadenfreude 152
sclera 134
selection 2–10, 18–35, 48–62, 136, 141, 150, 159–189, 216–218, 222, 236, 241, 290, 301, 325, 333
antagonistic 173
counter-gradient 173
cumulative 35
frequency-dependent 213
history of 25, 30, 158, 175, 272, 290
kin 220
natural 2–10, 18–35, 41, 48–62, 66, 73, 77, 81, 86, 93, 96, 100, 102, 104–105, 108, 111, 114, 116, 127, 152, 155–156, 159–189, 192, 194–199, 202–204, 209–211, 219, 232, 261–263, 266–269, 283, 285, 290, 296–297, 299, 303–312, 329–330
relaxed 223n7
runaway 53. *See also* evolution (autocatalytic)
sexual 53, 326
self 124, 237, 261, 265, 282
adjustment 29
assembly 29
awareness 265
feeding 49, 53, 222, 230, 327–328
interest 253
organization 271, 337
propulsion 108–110, 123
report 138n2
structuring 227
semantic colorization 48, 85, 88, 92, 96, 122, 283, 337
sequential hermaphroditism 177
serial homology. *See* homology, paralog
serotonin 314
sexual dimorphism. *See* dimorphism
sexual reproduction 210
sexual selection. *See* selection, sexual
Shuar 234, 245, 250
shout 272, 284. *See also* demon, model (pandemonium)
signal 59, 111, 143, 162–163, 194, 211, 278, 301, 312–314. *See also* cue, index
biochemical 21–24, 61–63, 195, 285, 314
detection. *See* detection (signal)
signaling cascade 21–24, 26, 40, 61–63, 80, 301
simple effect 289
Smarties task 131
snake 6–9, 105, 112, 116, 233–234, 269
detector. *See* detector (snake)
social cognition 106, 126, 141, 255–256, 328
social construction 104
social learning. *See* learning (social)
society of mind 280–281
somatic marker 285
somatosensory homunculus 316. *See also* cortex (somatosensory)
space 50, 214, 218, 336–337
conceptual 57
design 35, 59, 76, 100, 161–167, 172–174, 219, 297
developmental 50, 54–55
environment 54, 169, 172–174, 176
evolutionary 50, 54
fitness 172, 181
functional 301
genotype 50–52, 211, 336
hypothesis 101
information 50, 56–57, 60, 76
massively multidimensional 37–38, 162, 175, 184, 203–204,. 337. *See also* space (n-dimensional)
morpho- 51, 56
n-dimensional 51. *See also* space (massively multidimensional)
phenotype 50–54, 62, 160, 163–167, 172–174, 176, 203, 211, 297, 336
possibility 23, 27, 29–30, 33, 35–36, 39, 50–56, 89, 152, 165, 186–189, 211, 223, 246, 336
state 58, 65, 68

strategy 213
work-. *See* model (workspace)
specialization 3–7, 10–12, 26–30, 41, 45–46, 54, 66–67, 74–77, 90–91, 94–95, 97, 99, 101–102, 106, 111, 113–120, 123, 125–129, 133, 137, 145–146, 150–151, 157, 171, 191–196, 199, 202, 204,, 215, 221, 225, 229, 232, 239, 248, 252, 263–267, 270–272, 276–281, 286–288, 293–294, 299, 301, 310–319, 326, 329–331, 334–338. *See also* domain specificity
hierarchical 302–308
speech parsing. *See* parsing, speech
state 41, 46, 148, 201
attentional 201–202
goal 46, 238–239, 295
mental 41, 47, 89, 105, 116, 128–138, 141, 144–152, 172, 183, 186, 199, 202–203, 278, 317, 331
of affairs 148–150, 200–201
space. *See* space (state)
stimulus enhancement 43
subfunctionalization 301, 304. *See also* duplication and divergence, paralog
subtractivism 19, 30–31, 35, 48. *See also* additivism
superior temporal sulcus (STS) 146
system 2, 5, 18–20, 23, 31, 37–41, 57, 59n2, 60–61, 69 72–76,81, 85, 86, 92–97, 99, 105–107, 109–111, 114–118, 120–122, 125, 130, 140–142, 150–152, 215, 263–264, 269–281, 283–287, 290–293, 307–314, 331–338
behavior guidance 23
central / peripheral 265
complex 289
coordinate 30
conceptual 86–89, 103
decision-making 92, 101, 269, 293–294, 309
developmental 10–11, 29, 32–35, 53, 56, 60, 62, 125, 156–181, 182–205, 226, 290, 297–298, 315–317, 319, 324, 338
dual. *See* model (dual systems)
dynamical 212, 222, 242, 256–257, 275n4, 290, 326, 337
gene regulatory 32, 56, 217, 305, 313–314, 318
higher-level 103, 265–267, 271, 276
inference 47, 88–91, 100, 108, 122, 246, 278–279
information-processing 60, 103, 210–211, 267
input 268–269
interaction 6, 106, 129, 295–296, 330
learning 36–37, 45, 91–92, 96–97, 101–102, 211, 221, 227–228, 233–235, 248
lower-level 2, 126, 265–267, 269, 271
memory 74
module spawning 188–119, 309
motivational 87, 240, 252–255, 275
neural 61–63, 273, 302–304, 314
self-organizing 271
signaling 24
visual 65–68, 70, 80, 111–112, 119, 149, 262, 297
System 1. *See* model (dual systems)
System 2. *See* model (dual systems)

tag 48, 67, 85–86, 92–93, 101, 105, 108, 110, 115–116, 148–150, 275–279, 293–296
contextual 280
informational 107, 115, 274, 312
neural 285
representational 148, 306
semantic 294
template 6, 114–115, 126, 308–309
animate 114
animal 114
face 116–117, 278–280
innate 156–157
object 309–310
predator 233–234, 309
snake 234
spawning 114–115
spider 234
The Automated Set of Systems 265
theory of mind 41, 70n5, 128–130, 133, 138n2, 145–146, 149–151, 199–202, 211, 215–217, 222, 237, 239, 246, 255, 292, 300, 328, 331, 335. *See also* intersubjectivity, inference (intentional), mindreading

time-series analysis 203
timescale 76, 312
  cognitive 76, 312
  cultural 291
  developmental 76, 312
  evolutionary 76
  neural. *See* timescale (cognitive)
  real-time. *See* timescale (cognitive)
token 114–115, 125, 155, 182, 188, 191, 193, 195–197, 203, 215–216, 225, 236, 238, 244, 251, 301, 309–310, 318. *See also* type
tonic immobility 134, 310
tracking 59
  agent 202
  attention 202
  belief 132–133, 135, 141, 145, 147–148, 199–202, 300
  gaze 135
  identity 70–71
  kind property 71–72
  mental state 105, 131, 135, 137, 145
  motion 129
  object 68–75, 85–86
  property 71–75
  spatiotemporal 71–76, 88
  world-state 202
tradeoff 8, 20, 74, 158, 165n5, 167, 170, 195, 268, 270
  cost-benefit 136, 178
  decision-making 286
  flexibility-specialization 276
  quality-quantity 8, 112
  specialization-generalization 74, 76, 276
trait 10, 28, 53, 159, 168, 174–175, 177, 179, 183, 230, 298–300, 322. *See also* phenotype
  ancestral 325–326
  canalized. *See* canalization
  derived 46, 325–332. *See also* autapomorphy
  heritability of 162n4
  homologous 46, 300. *See also* homology
  in disequilibrium 78
  innate 156, 174
  novel 300–301
  plastic. *See* plasticity
transduction 2–3, 21, 267–268, 301
transmission 43, 53, 225–242. *See also* cultural transmission, bias
  biased 228–231, 235
  conformist 228–230
  cultural 53, 129, 210, 225–242, 246, 252, 328, 335
  information 59, 160, 210
  language 217
  mechanism. *See* mechanism (transmission)
  rule. *See* rule (transmission)
  signal 63
  social 43, 45, 232
  synaptic 315
triadic awareness 240. *See also* attention (shared)
truthiness 268
type 48, 114, 182, 191–193, 196, 212–213, 302–303, 312, 318. *See also* token

uncertainty 58, 198
unconsciousness 5, 128, 145n5, 266, 281–285
underdetermination 36–38. *See also* problem (frame), problem (opacity)
uniqueness. *See* human uniqueness
Universal Grammar 37, 215
universality 192, 324
updating 79, 95
  attitude 284
  Bayesian 173, 203
  deliciousness 93
  food 232
use it or lose it 18

validity 45, 111, 232–233
  cue 111
  ecological 45
  predictive 232–233
valuation 88, 96–99
variation 9, 32–35, 51–52, 62–63, 66, 76, 111, 117, 158–193, 202–203, 209, 218, 227, 236, 251, 254, 278, 323–326, 328
vasopression 314

ventral stream 74
veridicality 268
violation of expectation paradigm 71, 122.
*See also* looking time paradigm
visual agnosia 111, 118
Visual Word Form Area 307
visuospatial sketchpad 286

WEIRD society 324
wolf. *See* canid