

Reviews

Natural Minds

THOMAS POLGER

Cambridge, MA, MIT Press, 2004

xxvii + 294 pages, ISBN: 0262162210 (hbk); \$40.00

In *Natural Minds*, Thomas Polger joins a growing number of theorists who defend the mind-brain type-identity theory while casting doubt upon orthodox non-reductive varieties of functionalism in the philosophy of mind. Polger has written a fine book in a fast-paced style that covers a lot of ground. He discusses different kinds of multiple realizability and their support for functionalism (ch. 1), Kripke's modal argument against the identity theory and Levine's explanatory gap problem for physicalism (ch. 2), the many possible functionalist views about the mind (ch. 3), various notions of the realization relation (ch. 4), and problems for specific kinds of functional realization with respect to desiderata like objectivity, biological abstractness, and causal efficacy (ch. 5). Polger also criticizes arguments that might be used to support orthodox nonreductive functionalism that are based upon Shoemaker's fusion of causal-role functionalism and a causal theory of properties, Lycan's analysis of levelhood in terms of functional roles and their occupants, Putnam and Fodor's appeal to the methodological autonomy of psychology, as well as Machamer, Darden, and Carver's multi-level account of mechanistic explanation that entails neither reduction nor property identity (ch. 6). Polger also includes a discussion of Dennett's challenge to philosophical intuitions about zombies, which is extraneous to his main thesis but connected to matters of mechanistic explanation (ch. 7), and he concludes with a curious concession about one alleged counter-intuitive consequence of the identity theory—namely, a cell-assembly that has become dissociated from the individual would still possess sensations (ch. 8).

Much of Polger's book provides a convenient summary of the literature. For example, Polger presents several familiar strategies to defend the identity theory and settles upon Lewis and Kim's evolving proposals about species-specific properties coupled with Kim and Adams' observation that mere physical differences between conscious brains does not exclude neurophysical commonalities that might underwrite mind-brain property identities (ch. 1). Polger also presents the much-discussed family of problems regarding functionalism and the causal efficacy of mental properties (ch. 5). But he organizes the philosophical landscape in a useful way. Especially noteworthy is Polger's taxonomy of possible functionalist positions,

whose categories can be extended outside functionalist theory (ch. 3), as well as his scorecard of questions about consciousness and zombies, which brings order to that strange fantasyland where the dead walk and qualia dance then disappear (ch. 7). More central to his main thesis, Polger moves the debate over mind-brain identity forward with provocative arguments that deserve close attention. Yet I do not believe that Polger's *Natural Minds* is likely to change many minds. I will mention just two sample concerns.

My first concern is about Polger's charge that orthodox functionalists beg the question when they appeal to multiple realizability (MR). Polger begins by distinguishing four kinds of multiple realizability, including two that are often advanced together and directly implicated in his charge: "Standard MR" whereby "systems of indefinitely (perhaps infinitely) many physical compositions can be conscious," and "Radical MR" whereby "any (every) suitably organized system, regardless of its physical composition, can be conscious" (p. 6). Polger then says:

But if a variety of multiple realizability is to be the basis for an argument against the identity theory, then it will have to be a form of multiple realizability whose plausibility does not itself depend on the truth of functionalism . . . I contend that Standard MR and functionalism go together. Standard MR is not part of any argument against identity theory that is independent of substantial metaphysical claims. (p. 7)

If Standard MR is the consequence of prior metaphysical commitment, then it is question-begging with respect to the identity theory. (p. 8)

I dismiss as hogwash anyone's claim to have metaphysically neutral and prefunctionalist intuitions in favor of Radical MR. (p. 9)

In this series of remarks Polger appears to make two different complaints. One is the charge about "begging the question." The other is a statement that Standard and Radical MR are not "independent of substantial metaphysical claims." If Polger means that Standard and Radical MR are not independent of substantial *functionalist* metaphysical claims, then it is the charge about begging the question. But if Polger means that Standard and Radical MR are not independent of substantial metaphysical claims, *even if they are independent of functionalist metaphysical claims*, then he has made a different assertion that is far more difficult to assess because it involves the role of metaphysics, how background assumptions legitimately operate in dialectical exchanges, confirmational localism versus confirmational holism, and other things endemic to the human philosophical condition.

So consider the more straightforward charge about begging the question. Roughly speaking, people beg the question when their conclusion is assumed by one of their premises. Thus, Polger claims that orthodox functionalists beg the question because their version of functionalism is assumed by their premises about Standard or Radical MR. Yet there are many distinct and independent notions of functionalism, even among the orthodox functionalists, as Polger emphasizes later in the book

(ch. 3). Hence one must choose a specific kind of functionalism to play the part of the conclusion in the imagined supporting argument from Standard or Radical MR. For the sake of illustration, let it be *machine* functionalism. Does a premise stating that mental properties are subject to Standard or Radical MR assume machine functionalism? No. The phenomenon of multiple realizability is independent of machine functionalism, since it can be explained in *nonmachine*-functional ways. For example, Dennett (1991) and Rosenberg (2001) have argued that multiple realizability can arise by a process of evolution, i.e., by random variation and natural selection due to the benefits of neural plasticity. But evolution has nothing essential to do with machine functionalism.

Worse, as the evolutionary explanation suggests, multiple realizability can be explained in *nonfunctional* ways. For a different case, Batterman (2000) offers an explanation based upon the notion of UNIVERSALITY in physics, i.e., the method of finding similarities in behavior among physically diverse systems. Or again, I stated some time ago that multiple realizability might be explained by appeal to “anomalous realization” whereby the variability in question arises because of lawless behavior, or by appeal to “macro realization” whereby the variability in question arises because of processes that generate the macro from the micro, or by appeal to “functional realization” whereby the variability in question arises because of the fact that different physical types can occupy the same functional role (Endicott, 1989, pp. 213–214). Granted, the first two are mere schema for constructing nonfunctional explanations, but I assume that details could be supplied which do not presuppose the heterogeneous class of functionalist notions in the philosophy of mind. Thus, one can explain the micro–macro distinction by nonfunctional notions like the mereology of particulars, and then explain the lack of lawful coextension between their micro–macro types or properties that multiple realizability requires by a process of mereological recombination for instances of those types or properties. Crudely put, take a gold statue and exchange its pieces with the appropriately fitted iron parts and the result is an iron statue—the same type of macro shape now multiply realized by different types of micro-material constitutions. I also believe that these roughly Aristotelian-inspired ideas predate contemporary functionalist intuitions.

In fact, one can press the matter of nonfunctional multiple realizability by appreciating the implications of “physical multiple realization.” Like a specific degree of temperature or mean kinetic energy versus the many types and arrangements of micro particles that serve to realize it, higher-level physical properties can be multiply realized by lower-level physical properties (a fact, ironically, that identity theorists fondly point out; see Bickle, 1998, pp. 124–126). So unless Polger is willing to extend functionalism all the way down through the physical levels of reality, and he is not (pp. 186–190), he must concede that multiple realizability is both conceptually and factually independent of functionalist considerations. That is, functionalism and multiple realizability do not always “go together” (for a comprehensive survey of issues pertaining to multiple realizability, see Endicott, 2005).

My second concern is about Polger's general methodology relative to mind science. He announces at the beginning that *Natural Minds* is a "work of philosophy," more specifically, a project in "naturalized metaphysics" (p. xv). As such, it represents a traditional philosophical practice dominated by methods of conceptual clarification that apply only at the periphery of actual scientific theories, testing, and results. Of course the philosophical groundwork for theory construction is important. Moreover, Polger's task is a modest one that "depends a great deal on clearing the philosophical landscape and making it safe to be an identity theorist" (p. xxii). In other words, Polger only aims to establish that the identity theory is consistent with the empirical facts, not that it is best supported by the facts. Indeed, Polger is skeptical about the prospects for resolving the debate over functionalism versus the identity theory by appeal to current scientific evidence: "I doubt that the evidence is even suggestive in this respect, because I doubt that the evidence by itself can settle metaphysical questions such as that between functionalism and identity theory" (p. 24). Yet, I worry that Polger appears too cavalier about cognitive science by dismissing its many applications to important philosophical questions about the nature of mind.

Polger thinks that "conscious experience is the fundamental mental phenomenon" (p. xxi), contrary to an almost complete scientific consensus that consciousness is a small but significant part of human mental processes. Even more surprising, Polger defends the mind-brain type-identity theory for conscious experience without advancing a single scientific theory of consciousness and without postulating a single neural correlate for consciousness. Or again, when addressing Kripke's infamous modal intuitions about pain without its brain correlates, Polger dismisses them by maintaining that people have no firm grasp of the identity conditions for either mental states or neural states (pp. 49–51). For example, he says: "We do not know how to individuate brain states, properties, processes events, and so forth. Not only do we not know how to individuate these things, we don't really even have a clue what such things are" (p. 51). Not a clue? The scientists are lost. And more candidly, from a personal perspective:

I do not know how to thoroughly defend the negative existential claim I am making: that as a matter of empirical fact at this time, we do not know the identity conditions for brain states, properties, processes, events, and such. How does one defend the claim that some putative bit of knowledge is not now had? The best I can do is to consider the only candidate I know of for our knowledge about brain states: brain-imaging studies. (p. 52)

Polger then discusses some results from MRI studies and concludes that they are insufficient to supply the needed information about brain-state individuation (pp. 53–57). But even if Polger is right about MRIs, this is but *one* source of scientific knowledge that must be considered together with the results from other technologies, data supplied by many different kinds of studies in brain deficits from clinical neuropsychology, research in brain development and evolution, and the myriad of behavioral data, all of which must then be placed in reflective equilibrium with our

best theories of neurophysiology, neurocomputation, and neuropsychology, in conjunction with our best theories of cognitive psychology, psycholinguistics, and computational modeling. In short, Polger's defense of his negative existential claim requires a Herculean task that he has only begun to perform. But these and other problems notwithstanding, Polger's *Natural Minds* is a philosophical text well worth reading.

References

- Batterman, R. (2000). Multiple realizability and universality. *British Journal for the Philosophy of Science*, 51, 115–145.
- Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.
- Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown, and Co.
- Endicott, R. (1989). On physical multiple realization. *Pacific Philosophical Quarterly*, 70, 212–224.
- Endicott, R. (2005). Multiple realizability. In D. Borchert (Ed.), *The encyclopedia of philosophy* (2nd ed., Vol. 6, pp. 427–432). New York: Macmillan Reference USA.
- Rosenberg, A. (2001). On multiple realization and the special sciences. *Journal of Philosophy*, 98, 365–373.

RONALD ENDICOTT
Department of Philosophy & Religion
Director for Cognitive Science
North Carolina State University, USA
Email: rpendico@ncsu.edu

Understanding Phenomenal Consciousness

WILLIAM S. ROBINSON

Cambridge, England: Cambridge University Press, 2004

xii + 264 pages, ISBN: 0521834635 (hbk); \$65.00

There's been much recent debate around the question of whether phenomenal properties can be accounted for in terms of physical properties. Property dualism answers "no" to this question, and is generally accompanied by the epiphenomenalist claim that phenomenal properties are irrelevant to a causal account of behavior. The publication of Robinson's book on phenomenal consciousness is welcome, as it provides a comprehensive property dualist and epiphenomenalist picture that approaches the issues of well-worn controversies in a different manner. First, Robinson is sensitive to the fact that much of the discussion concerns the move from the observation that phenomenal properties are not explainable in physical terms to ontological conclusions that set them apart from the physical domain, and much of

the book focuses upon strengthening the argument against materialism on this point in a novel way. Second, Robinson is also aware of the accusation that property dualism is obstructing the progress of a scientific understanding of consciousness. An important theme running through his book is thus the distinction between scientific progress and the materialist paradigm. Additionally, Robinson displays an impressive knowledge of the psychological research into consciousness, and uses it to give empirical backing to his argument.

Robinson's book is divided in two parts. The first develops the argument for a property dualistic ontological position: "Qualitative Event Realism" (QER). The second part examines how an understanding of the physical causes of phenomenal consciousness could be obtained, so as to pave the way for a broader non-materialistic but still naturalistic theory of consciousness. The way Robinson presents the problem of phenomenal consciousness is concise and clear. Focusing upon visual perception, he gives a full account of the physical processes involved in the perception of a red apple, and asks how color comes into the story (p. 8). At the outset, the answer that it's through the perceived apple's having the property "red" is discounted as generally unsatisfactory. This move may seem a bit quick, and it is only later that Robinson properly addresses a closely related position which views qualitative events as illusory (p. 147). He locates the impasse at the point where the "skeptic," though agreeing that there's phenomenal consciousness, points to the gap between intersubjective agreement and the scientific (third-person) objectivity as sufficient to invalidate realism about phenomenal qualities. Robinson rightly asks whether such skepticism is making illegitimate demands for third-person proof. But this issue strategically emerges after eight chapters where the author has been engaging with the reader in constant reflection about how to account for phenomenal properties on different understandings of consciousness, so that the skeptic's claims lose much of the force they may have possessed.

The first two chapters lead to the precise formulation of QER while addressing certain basic objections to this realism about phenomenal properties. Robinson concurs with Wittgenstein (p. 20) that it's through intersubjective agreement that people learn to use words such as 'red' (which refers to the phenomenal property). This does not bar them from using such words to refer to their experiences, whereby they may occasionally make errors about the correct attributions of colors. One particular view that Wittgenstein's reflections were particularly directed against is the notion that, in afterimaging, one must distinguish between the afterimage and the experience of the afterimage, thus leading to realism about sense data. Robinson focuses upon Moore's arguments (pp. 23–25) and forcefully shows the shortcomings of any attempt to reify phenomenal consciousness: "QER is not a substance theory." Rather, instantiations of phenomenal properties are ways of being phenomenally conscious. The pair constituted by such a phenomenal quality and the duration of the experience constitutes the kind of event that QER is about.

The third chapter contains the core argument for QER. Interestingly, this is developed on the basis of Descartes's fallacious argument for the non-material nature

of the “I.” Robinson adapts this argument to a subject *S* and properties *F* (e.g., phenomenal property) and *G* (e.g., physical property) as follows. Assume that *F* is an appearance-constituting property, i.e., that something now appears *F* to *S*. Assume *G* is not such a property. The conclusion that *S* does not now have an appearance that is *G* turns on the following additional premise: “appearances actually *have* the properties by which they are *constituted*, but they *do not have* any non-relational or differentiating properties other than those by which they are constituted” (p. 39—with a confusing typographical error corrected). The limitations in this premise are required because there are relational properties (e.g., occurrence at a certain time) and non-differentiating properties (e.g., being temporal) (p. 40).

The conclusion is, however, insufficiently grounded to prove property dualism, for it relies upon the additional premise that a mode of appearance has no non-presented properties. Unlike Chalmers (1996), Robinson stresses the fact one cannot disprove a materialist who claims that, although we cannot explain how or why, “red” just is a physical property. This negative stress is however counterbalanced by an important positive metaphilosophical point: it is not the rejection of materialism, but rather adherence to an empty materialism which accepts as factual what cannot be proven, that is anti-science (p. 46).

Of course, most forms of materialism also reject this empty assertion of identity. An immediate way of providing such an explanation consists in providing an account of why there’s identity although we cannot know it. The first form of this approach, which is very popular, consists in showing that the problem lies in the different ways that physical and phenomenal *concepts* are acquired (Hill, 1991). The second makes the strong claim that our minds are limited in such a way that we aren’t able to understand how these properties can be identical (McGinn, 1991). In both cases, Robinson convincingly brings out the lack of explanatory power characterizing the approaches (pp. 46–52).

The rest of part two is then devoted to more complex attempts to provide an account of the problematic assertion of identity. Robinson starts his investigation with two chapters on representationalism—the theory that what has to be added to the account of the perception of a red apple in terms of the property of physical redness are representations of physically red colored things, and such representations are assumed to be amenable to a physicalist account (p. 55). For Robinson, the test of adequacy of such a theory is whether it’s able to address the perception/thought (P/T) problem, i.e., give a representationalist account of the intrinsic difference between perception (where we encounter phenomenal consciousness) and thought (which has no phenomenal quality). One representationalist proposal involves saying that perception involves having a representation of oneself seeing a red apple (p. 61). But to say that if I were having this thought, I would also be having a phenomenal experience involving the color “red,” amounts to saying that this phenomenal consciousness is additional to the representation, hence remains unexplained. Another representationalist view is that the content of the representation involved

in phenomenal consciousness is nonconceptual (p. 67). But this only provides an extrinsic characterization of the P/T difference, i.e. in terms of what subjects who have such representations can do with them. Finally, adverbialism is considered. Robinson argues (somewhat too quickly) that the only viable form of adverbialism involves representations that might not be naturalistically reducible and this theory is not obviously distinct from experiential realism (p. 71).

The next chapter presents a thorough analysis of the claim that “experiences are transparent representers of qualities of things” (p. 73) that exhibits the inadequacies of such an approach, while the following two chapters tackle higher-order theories (HOTs), respectively, of consciousness as higher-order thought and higher-order monitoring. Here, much of the crucial argument is encapsulated in the first ten pages. Because of the lengths of these chapters, it would have been helpful to indicate to the reader what could be skipped during a first reading. Essentially, Robinson challenges HOTs to “explain why adding sensory representations to thoughts about sensory representations should result in consciousness” (p. 128). Robinson recognizes that he cannot rule out the idea of unconscious phenomenal qualities, which could help the HOT case. However, without an account of how such qualities could contribute to consciousness, any HOT account of phenomenal qualities is incomplete and there are no prospects of any completion here.

In his comparatively and inexplicably short chapter on functionalism, Robinson shows that only microfunctionalism could be of explanatory value, through an account of the fine internal structure of the mechanisms in the brain that are associated with phenomenal consciousness. In discussing this approach, Robinson claims that there’s no significant difference between the search for the functions whose satisfaction suffices for phenomenal consciousness, and QER’s investigation into the causes of phenomenal consciousness. He adds that defining a phenomenal quality as a function that can in principle be realized in different media is not a form of materialism, so that there is in no obstacle to viewing QER as a type of functionalism (p. 140). The discussion of this important issue is not given enough prominence: its relation to Chalmers’s dualism is not clarified. While the latter’s ontology separates phenomenal from functional/physical properties, for Robinson, some functional properties of physical substances apparently shift to the first class of the divide. A further issue which is left unsatisfactorily open is that of whether strong supervenience could describe the relation of the phenomenal to the physical (p. 141). For surely, if QER is a form of property dualism, then strong supervenience involving logical necessity is out of the question. Although Robinson wishes to avoid discussion of modal issues, these issues cannot be totally bypassed in the light of the importance of their importance in current debates.

Part one concludes with a defense of the epiphenomenalism. Robinson draws upon empirical evidence from a related problem of the efficacy of our will in action to suggest that we are not in a good position to distinguish causation from common effects (p. 163). This is useful as part of a strategy to neutralize our antiphenomenalist intuitions that tell us that the phenomenal effects of physical

events are also causes of our behavior. Conceptual arguments are also provided to defuse antiphenomenalist concerns. In part two, Robinson puts forward his proposal for a naturalistic understanding that would integrate both the neural and the phenomenal levels. Possible approaches based upon the intrinsic nature of the physical or quantum mechanics are considered and rejected for not offering sufficient promise. The final two chapters consider the pros and cons of Robinson's proposal, namely that it is neural patterns that hold the key to understanding the cause of phenomenal events (p. 213). There are some interesting arguments here, but some of the discussion requires going along with speculative choices that are more or less well motivated. This puts heavy demands on the keenness of the reader who is asked to consider problems that *might* emerge *should* certain empirical facts turn out to be true.

Robinson's writing throughout the book is clear, and the arguments are concise. Although the first chapters are accessible to a wide audience, the book later suffers from not infrequent appeals to specialized knowledge: this may be in philosophy (e.g., externalism, Twin Earth, p. 154), or science (e.g., glial cells, p. 158; statistical notions such as the standard deviation and standard error, pp. 163–64; microtubules, p. 193). In general, the nature of the subject would seem to require that the author address a variety of types of readers. Since footnotes cannot do justice to all the details that are of interest, a different font could have been used for any discussion or piece of information liable to be of interest to a more specialized audience. Aside from the problems of the treatment of functionalism and modal issues, this book presents a powerful case for property dualism. It does this, both by arguing for the view and by showing the limitations of the resources of materialism in tackling the issue. Its great strength is a clarity and honesty about how far the argument against materialism is able to go. And in most cases, it would seem to be quite far enough to be persuasive.

References

- Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford, England: Oxford University Press.
- Hill, C. (1991). *Sensations: A defense of type materialism*. Cambridge, England: Cambridge University Press.
- McGinn, C. (1991). *The problem of consciousness*. Oxford, England: Blackwell.

CHRIS J. ONOF
Birkbeck College
University of London
London, England
Email: c.onof@imperial.ac.uk

Seeing and Visualizing: It's Not What You Think

ZENON W. PYLYSHYN

Cambridge, MA: MIT Press, 2004

581 pages, ISBN: 0262162172 (hbk); \$50.00

In *Seeing and Visualizing*, Zenon Pylyshyn presents a comprehensive discussion of two important and related problems in cognitive science: the relation between seeing and thinking and the nature of mental imagery. The book is an excellent example of how to construct a thesis drawing on material from related disciplines including psychology, philosophy, vision science, neuroscience and neuropsychology.

Although he does provide an overview of the debates in the area, the main motivation for the work is the presentation of Pylyshyn's own theories. The connected theses he presents are that seeing is a separate function from cognition which does not involve the construction of an internal pictorial display, and that the use of imagistic reasoning does not depend on a system of internal representations that are themselves intrinsically pictorial in nature. He argues that in order to reach a scientific understanding of seeing and visualizing, it is necessary to question "the view from within"—the misleading subjective perspective that suggests we literally construct an internal pictorial representation when we see and use imagery when we visualize. He both provides a substantial critical overview of the relevant literature, as well as goes some way toward offering alternative explanations for the persistence of our intuition of an internal image that is perceived by the mind's eye.

The book can be divided into two main sections, which respectively deal with problems of seeing and problems of visualizing. Chapters 1–5 are concerned with questions regarding the relation between seeing and thinking, and the nature of visual content. These chapters are used by Pylyshyn to argue for the modularity of vision and to establish his theory of "visual indexes"—mechanisms that operate prior to focal attention to bind parts of visual representations to their referents in a similar way to indexical terms such as 'this' or 'that'. These chapters centre on the presentation and interpretation of material from vision science, cognitive neuroscience, and cognitive psychology.

Chapter 1 introduces some of the unique problems accompanying seeing—in particular, the misleading nature of the phenomenology connected to it. Pylyshyn contrasts the picture created by empirical evidence with the sensation of consulting a detailed, stable picture of the world that is present to our introspection. The problem is then how to explain this phenomenology given that the input that we actually receive is neither detailed nor stable. He amasses a substantial and convincing body of evidence against the view that we create an inner picture, and foreshadows the idea of visual indexes by suggesting that many of the problems faced by traditional views stem from the fact that they do not provide a means of achieving indexical reference.

Chapters 2–3 deal with the relation between vision and cognition, and the nature of the architecture of the visual system. Pylyshyn raises some problems concerning the terminology involved in discussing 'vision', and restricts the term to the

operation of the early visual system. He considers vision to be modular in the Fodorian sense of being informationally encapsulated in respect to cognition. This means that it is “cognitively impenetrable” and cannot be affected by our beliefs, desires or knowledge. This leaves two problems: first, the problem of how to explain the evidence provided for the continuity of vision and cognition, and second, the problem of how to account for cases where the operation of the visual system appears to be intelligent. Pylyshyn considers much evidence presented for the continuity of vision and cognition, including evidence from neuroscience that suggests the influence of “higher” cognitive levels on the operation of the sensory level, and the successful use of knowledge-based systems to solve visual processing problems in robotic vision. However, he argues that the evidence is mostly a result of a terminological confusion in the use of ‘vision’, or that it does not fulfill the criteria of cognitive penetrability. Here, his restricted definition of ‘vision’ may leave some feeling that he has evaded certain problems; however, Pylyshyn convincingly argues that providing such a definition is a necessary step. He then presents several reasons for questioning the continuity thesis, including Fodor’s work on illusions, information regarding the differing principles followed by perception and reasoning (such as the fact that principles of visual organization are not rational), neuroscientific evidence, and evidence regarding the dissociation of visual and cognitive functions in the brain.

In order to address the problem posed by “intelligence” in vision, Pylyshyn turns to hardwired or architectural constraints of the visual system that might provide alternative explanations. In particular, he discusses architectural “natural” constraints that he suggests govern the interpretation of input according to embodied principles that offer a likely—though fallible—construal of the distal cause of input received. The fallibility of these interpretations is used to explain some well-known visual illusions. Pylyshyn also accepts that there exist both pre-visual cognitive influences (such as attentional selection) and post-visual influences (such as selecting between two competing interpretations of the same stimulus), however these do not threaten the thesis of the cognitive impenetrability of vision.

Focal attention is of particular interest as it forms the boundary between vision and cognition, and is discussed in detail in chapter 4. The main question that Pylyshyn addresses is the nature of the subject of this attention—whether focal attention works on locations, objects or some other properties of the world. He argues that focal attention is directed at visual *objects*, notwithstanding considerable literature in the field that suggests that it is directed toward locations. Furthermore, he argues that it is directed using a “visual index” or FINST, which is deployed prior to focal attention. This theory forms one of Pylyshyn’s original contributions to the debate.

Chapter 5 ties together the material from the previous chapters in order to more fully develop the theory of visual indexes and describe it in relation to both pictorial and descriptivist theories. Pylyshyn argues that objects are indexed directly (demonstratively referring to ‘this’ or ‘that’) and nonconceptually as opposed to

by describing their features or location. Visual indexes bind representations to the things they represent and keep track of objects identified through their causal history in the face of the constant visual changes that they undergo. While Pylyshyn offers some empirical support for the visual-index theory, he also indicates areas for future research, and the theory is not an established one.

Chapters 6–8 mark a shift in focus from the problems of seeing to questions regarding the use of mental imagery. Pylyshyn supports Escher's view that "a mental image is something completely different from a visual image" (p. 332), and argues that there is nothing inherently pictorial about the representations that accompany tasks employing mental imagery. Visual tools can help us to think, but they are not the stuff of thought. While still drawing on evidence from cognitive psychology and neuroscience, the nature of the questions discussed in these chapters makes them more speculative and philosophical in tone. Chapter 6 involves a discussion of empirical evidence that is suggestive of the imagistic nature of representations, while chapter 7 is more concerned with analysis of the coherence of the position that we literally think using visual images. These chapters offer a much-needed clarification of the issues in the mental imagery debate. He differentiates between the form and content of representations, and argues that for mental imagery to literally be pictorial, the *form* of representations must have imagistic properties such as space as part of their intrinsic nature. As a means of testing this, he uses the idea of cognitive penetrability as a "litmus test" for identifying the properties of representations that are part of their intrinsic nature. Many of the visual effects that we experience are cognitively penetrable. One of the alternative explanations for their persistence is that we draw on our tacit knowledge of how the real, spatial world works in order to construct our representations, although the possibility that imagery utilizes a special (although non-imagistic) form of information processing allows for the explanation of some effects. Pylyshyn also uses this possibility to explain the fact that many imagery effects are experienced by the congenitally blind.

Pylyshyn considers both the claim that there is an actual spatial display in the brain—which he dismisses as largely implausible—as well as the more widely accepted theory that mental imagery uses a "functional" space. He argues that functional space has no *intrinsically* spatial properties. It is rather a set of externally imposed restrictions on information that are not bound by the laws that restrict physical space, and so can be accommodated by propositional representations. Alternative explanations for the persistence of spatial effects include our use of external space and our own sense of proprioception as an aid to anchoring spatial information in some reasoning tasks, as well as a difference in the content of representations deployed.

In chapter 8, Pylyshyn considers in more detail the nature of the form of representations and how we might use visual displays to help us reason. He argues that considerations such as explaining the productivity and systematicity of thought lead to an acceptance of a "language of thought" that differs from both pictures and natural languages in certain key respects. While Pylyshyn presents this as a relatively uncontentious restriction on mental content, it is not clear that

his arguments to this point have established this conclusion, and an acknowledgment of the ongoing debate in this area would have been useful. Pylyshyn also discusses the various ways in which visual displays (e.g., diagrams) might be used in reasoning, and the status that they carry in terms of providing proofs. He argues that although they have no privileged status in the sense of having the same form as internal representations, they can help us to make explicit what is known implicitly and aid us in directing our attention to certain properties of the world.

Pylyshyn's book is not intended as a history of ideas so much as a presentation of the current state of the field and a background to his own theories. As such, some may feel that the characterization of particular positions is impoverished. Those after a more balanced view of, for example, the imagery debate might wish to supplement this book with a reading of material from the "other side" (e.g., Kosslyn, 1994; Tye, 1991). There are some other minor problems that attend the broad scope of the book. Detail sometimes overwhelms the clarity of the argument in the earlier chapters, and some of the philosophical and other issues raised could have been given more extended treatments in places. However, these are minor problems that never seriously detract from the work. The broad range of the book is required to do justice to the discipline of cognitive science and to provide the insights that are offered, and such issues are inevitable.

While the claims made in the book may appear bold, they are probably most contentious for those working in the areas of psychology that still operate under the assumptions that he questions. They are rendered less radical than they might otherwise be by the restricted definition of 'vision' that Pylyshyn adopts, and also by the model of explanation that he uses in which the explanatory burden of the representational system is less than that shouldered by some other representationalist theories. Interestingly, he works some of the insights of nonrepresentational theories and Gibsonian theories into his position. His position is for the most part a persuasively argued alternative to the critiqued views, although—as noted earlier—it is not clear that some of the conclusions drawn in the latter part of the book are as well supported as earlier material.

Aside from the material presented, the book also introduces many further questions and areas of research. In particular, the methodological issues discussed in the book raise questions regarding the possibility of investigating the nature of consciousness and how this is tied to our understanding of issues surrounding sensory experiences.

Although dense, Pylyshyn's book is a rewarding and detailed discussion of some central problems in cognitive science. The nature of the subject matter means that it is of interest to anyone working in cognitive science, psychology and philosophy, and not just those with a specialized interest in vision. While not everyone will be interested in the detail presented, it should satisfy those with a background in many disciplines, providing insight into related areas and a new perspective on some familiar problems.

References

- Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.
- Tye, M. (1991). *The imagery debate*. Cambridge, MA: MIT Press.

ALICIA CORAM
Department of Philosophy
Monash University, Clayton Campus
Wellington Rd, Clayton
Victoria, 3800, Australia
Email: aliciacoram@yahoo.com.au

The Oxford Handbook of Rationality

ALFRED R. MELE & PIERS RAWLING (Eds.)

Oxford, England: Oxford University Press, 2004

xii + 477 pages, ISBN: 0195145399 (hbk); \$74.00

Rationality has long been a central topic in philosophy, crossing standard divisions such as epistemology, action theory, and ethics. Since the mid-20th century, it has also become increasingly interdisciplinary, attracting attention from researchers of economics, psychology, law, education, and even neuroscience. *The Oxford Handbook of Rationality* provides a timely, authoritative, and state-of-the-art survey of current philosophical research in this vibrant and fertile area. Mele & Rawling have undertaken the impressive task of assembling 22 newly-commissioned chapters by a list of outstanding philosophers. The result is the first comprehensive and essential volume of its kind.

This book consists of two main parts. The first examines the nature of rationality from various perspectives; the second explores the relation of rationality to other specific domains of inquiry—including psychology, gender, personhood, language, science, economics, law, and evolution. The volume opens with a general introduction, in which the editors sketch the theoretical terrain and situate the collection. Customarily, the domain of rationality is divided into two subcategories: theoretical and practical. Whereas theoretical or epistemic rationality concerns what it is rational to believe, practical rationality concerns reasoning about and justification of decisions, motivations, and actions.

Three of the 13 chapters in Part 1 primarily focus on theoretical rationality. In “Theoretical rationality: Its sources, structures, and scope,” Robert Audi discusses the essential sources (e.g., perception, memory, consciousness, reason and reasoning, and testimony) of our rational beliefs, and examines the role of

coherence (as an internal relation among beliefs and concepts) in epistemic justification. For Audi, properties and principles of theoretical rationality are embodied in one's structure of cognition, which constrain the responsiveness to experience and the processes of belief-formation and -change. Audi also explores the scope of theoretical rationality and the criteria of being theoretical rational persons. Gilbert Harman's "Practical aspects of theoretical reasoning" discusses how and to what extent practical reasons are relevant to theoretical reasoning. People are resource-limited agents, and reasoning about one issue inhibits considering another. So, practical considerations are relevant to how much cognitive resources and effort to devote to a given inquiry, and when to terminate one. Harman also examines the role of simplicity and epistemic conservatism in theoretical reasoning, and their practical justification. James Joyce's chapter "Bayesianism" deals primarily with Bayesian epistemology, which requires that an adequate epistemology must recognize that beliefs come in varying degrees of strength, and which seeks to replace the categorical concept of BELIEF as an all-or-nothing attitude of accepting a proposition as true with a graded understanding of belief as level of confidence. Joyce explores the Bayesian requirement of probabilistic consistency, explicates Bayesian confirmation theory and the Bayesian theory of learning vis-à-vis rational belief change, addresses the most persistent objection to Bayesianism that it engenders an untenable subjectivism, and concludes that "Bayesianism remains without peer as a theory of epistemic reasons and reasoning" (p. 153).

Most of the essays in Part I focus on various aspects of practical rationality. One of the foci is the relations between motivation, reason, and rationality. In "Procedural and substantive practical rationality," Brad Hooker and Bard Streumer examine the debate between what they call "proceduralists" and "substantivists." According to proceduralists, agents can have reasons to have a desire only if they can rationally reach this desire from the beliefs and desires that they have. According to substantivists, there is also a different, substantive kind of practical rationality besides procedural rationality: agents can have reasons to have a desire whether or not they can rationally reach this desire from the beliefs and desires that they have. Hooker and Streumer examine the main arguments for both views, outline substantivists' criticism of proceduralism, and discuss the possibility of being a proceduralist about practical rationality while a substantivist about practical reason. Beginning with an often-quoted passage from Hume's *A Treatise of Human Nature*, which starts with the claim that "It's not contrary to reason to prefer the destruction of the whole world to the scratching of my finger," Michael Smith explains how Hume was led to his radical conclusion in "Humean rationality." For Smith, the explanation lies in Hume's view that the concepts of REASON and RATIONALITY are best explained by reference to their relations in the theoretical domain—specifically, deductive reasoning. Smith considers how we might avoid Hume's radical conclusion by liberating our understandings of the terms 'reason' and 'rationality', but still preserve a sensible view of Humean rationality, with some concession to Kantians. Onora O'Neill's chapter, "Kant: Rationality as practical reason," is a highly-readable exposition of Kant's attempt to account for practical reason that offers unconditional

reasons for action and provides the basis for a reasoned account of human duties. Unconditional practical reasons are those not based on arbitrarily chosen ends. According to O'Neill, Kant's proposal was that what makes a practical reason unconditional is its universal recognizability: a practical reason is an unconditional one, insofar as any rational audience can understand it as a reason for action. In "Duty, rationality, and practical reasons," David McNaughton and Piers Rawling present a view in which duty, rationality, and practical reasons are weakly connected. They take practical reasons not as motivating or explanatory psychological states, but as non-normative facts (e.g., the fact that the rubbish bin is full), as reasons for people to do something (e.g., take the rubbish out). Accordingly, rationality is just a matter of consistency, and duty is neither a matter of rationality nor of practical reason. They criticize various forms of internalism—especially Bernard Williams's—and propose a view of duty that is neither purely subjective nor purely objective.

Another cluster of essays in Part 1 focuses on the nature of practical reasoning, decision making, and rational choice—especially in a social context. James Dreier's "Decision theory and morality," shows how decision theory's formal apparatus is connected to abstract issues in moral theory. After explaining how to think about the concept of UTILITY and emphasizing that decision theory does not assume or insist that all rational agents act in their own self-interest, he discusses decision theory's contributions to social contract theory, emphasizing David Gauthier's rationalist contractualism. He also considers a reinterpretation of the formal theory with the implication that utility might represent goodness rather than preference, and discusses how Harsanyi's theorem provides an illuminating argument for utilitarianism. Game theory aims to understand situations in which decision makers interact. In "Rationality and game theory," Cristina Bicchieri focuses on the issue of whether rational choice theory is an adequate foundation for game theory in the context of noncooperative games, and examines some assumptions about individuals' mutual knowledge and beliefs related to practical rationality and whether they are sufficient to lead players to coordinate upon mutually-acceptable outcomes. Patricia Greenspan's chapter, "Practical reasoning and emotion," discusses emotion as an element of practical rationality. The currently-dominant approach (often referred to as "cognitivism" or "judgmentalism") rests on assigning emotions an evaluative content, which can in turn bias practical reasoning and decision making. Greenspan's essay contains a nice survey of judgmentalism and its variants, as well as its major alternatives—especially the causal/historical approach to emotions. Edward McClennen's "The rationality of being guided by rules," addresses a dilemma about whether it is rational to be guided by rules to which one has made a commitment: if there is a better option based on the balance of reasons, then it seems irrational to follow the rule; but if there is no better option, then the rule appears redundant. McClennen argues that this dilemma is resolved by revising the standard account of practical reasoning to accord with the prescriptions of a resolute choice model, in which a rational agent can choose to constrain future choices to what one has decided in advance, to a prior commitment to a rule.

The last two chapters in Part 1 focus on some deficiencies and puzzles of both theoretical and practical rationality. In “Motivated irrationality,” Alfred Mele examines two major kinds of irrationality: akratic action (i.e., action exhibiting so-called weakness of will or deficient self-control) and motivationally-biased belief, including self-deception. Roy Sorensen’s “Paradoxes of rationality” discusses more than two dozen paradoxes of theoretical and practical rationality, which appear to contradict some highly plausible principles (e.g., the principles of charity and of maximizing expected utility, the transitivity of preferences).

Part 2 of this volume consists of 8 chapters addressing rationality’s role in and relation to other specific domains of inquiry. It opens with Richard Samuels and Stephen Stich’s “Rationality and psychology.” Since the 1960s, psychologists have devoted a great deal of attention to human reasoning and decision making, collecting numerous experimental findings suggesting that human reasoning is—in important respects—normatively problematic or irrational. Samuels and Stich first present a brief sketch of some of these disturbing findings—including the Wason selection task and some key results from the “heuristics and biases” tradition. Then, they discuss the pessimistic interpretations of the findings that, in making judgments and decisions, ordinary people rely on heuristics rather than appropriate norms of rationality; these interpretations have been recently challenged by evolutionary psychologists, who hold varieties of more optimistic views about ordinary peoples’ rationality. Samuels and Stich, however, argue that—given the entire body of findings on human reasoning—the right position should be a “middle way” between the heuristic and biases tradition’s pessimism and the optimism of evolutionary psychologists: “People do make serious and systematic errors on many reasoning tasks, but they also perform quite well on many others. The heuristic and biases tradition has focused on the former cases, while evolutionary psychologists have focused on the latter” (p. 296). They further suggest that ordinary people’s reasoning and decision making are subserved by two quite different sorts of systems: one is fast, holistic, automatic, largely unconscious, and requires relatively little cognitive capacity; the other is relatively slow, rule based, more readily controlled, and requires significantly more cognitive capacity. Kirk Ludwig’s essay, “Rationality, language, and the principle of charity,” deals with the relations between language, thought, and rationality—especially the role and status of assumptions about rationality in interpreting others’ speeches and assigning contents to their psychological attitudes. Ludwig’s discussion is organized around three questions: (i) *what is the relation between rationality and thought?*, (ii) *what is the relation between rationality and language?*, and (iii) *what is the relation between thought and language?* Ludwig concludes that some large degree of rationality is required for both thought and language, as language requires thought but not vice-versa. Paul Thagard’s “Rationality and science” provides a review and assessment of central aspects of rationality in science. It deals first with the traditional epistemic question: *what is the nature of the reasoning by which individual scientists accept and reject conflicting hypotheses?* Then it discusses the natures of practical reason and of group rationality in science. As for the question *Are scientists in fact rational?*, Thagard suggests that

while “models based on formal logic and probability theory have tended to be so remote from scientific practice that they encourage the inference that scientists are irrational, . . . psychologically realistic models based on explanatory and emotional coherence, along with socially realistic models of consensus, can help to illuminate the often impressive rationality of the enterprise of science” (p. 379). Paul Weirich’s “Economic rationality” examines three views of rationality that dominate recent economic theorizing—namely, rationality as self-interest maximization, as utility maximization, and as bounded rationality. After reviewing criticisms of each, Weirich defends a refined version of the principle to maximize utility that adopts a broad interpretation of utility. Claire Finkelstein’s chapter, “Legal theory and the rational actor,” focuses on the normative thesis that the legal system should serve the goal of maximizing social welfare. Law and economics usually rely on descriptive theses about human nature—something which might be called “rational actor psychology”—and legal economists tend to accept utilitarian theories of value. Finkelstein shows that not only is rational actor psychology not inherently linked to utilitarian theories of value, the two are actually in tension. The tension stems from the fact that when we maximize social welfare, we do not necessarily maximize the welfare of each individual member of society. Finkelstein then argues that rational actor psychology more naturally leads to contractarian rather than utilitarian theories of legal rules.

Feminists have focused on reason and rationality ever since the emergence of modern feminist thinking, given that women’s previously-assumed deficiency in rationality has been used to justify their exclusion from full ethical standing as well as participation in education, politics, and the professions. In “Gender and rationality,” Karen Jones explores three major feminist stances toward gender and rationality: (i) the “classical feminist” stance, according to which what needs to be challenged are not available norms and ideals of rationality, but rather the supposition that women are unable to meet them; (ii) the “different voice” stance, which challenges available conceptions of rationality as either incomplete or accorded an inflated importance; and (iii) the “strong critical” stance, which finds fault with the norms and ideals themselves. The chief focus is the third critical camp and Jones’s chapter provides a sensitive survey and discussion of the arguments of some leading contemporary feminist thinkers. Carol Rovane’s “Rationality and persons” examines several related themes about our understanding of persons: unlike thermostats and pigeons, persons are not merely rational, but possess full reflective rationality; there is a single overarching normative requirement that rationality places on persons, which is to achieve overall rational unity with themselves; beings who possess full reflective rationality can enter into distinctively interpersonal relations, which involve efforts at rational influence from within the space of reasons; etc. Finally, in “Rationality and evolution,” Peter Danielson shows how game theory, the abstract theory of strategic interaction, supports powerful new insights into the relation between rationality and evolution. The striking claim is that rationality and evolution are isomorphic optimizing processes. Danielson sketches the main concepts and variants of evolutionary game theory, outlining how the differences between evolution and

rationality illuminate some crucial problem in the theory of rational choice. With special attention to the role of models and simulations, he proposes some low-level models that combine the two. Danielson concludes with a survey of the normative significance of unifying evolutionary game theory and rationality research, plus some speculation about human rationality's evolution.

As a whole, the *Handbook* provides an engaging and accessible survey of current philosophical studies of rationality, and a useful up-to-date roadmap of state-of-the-art thinking on this enduring topic. Of course, the coverage of the *Handbook* might have been more comprehensive. For instance, in Part 1, a separate essay on collective rationality—a topic taken up in a number of different chapters, but which deserves concentrated treatment on its own—would have been complementary. Likewise, one might have expected to see chapters relating rationality to education, culture, or postmodernity in Part 2. Overall, this is an authoritative sourcebook. Whereas each chapter can be read on its own and the references within each essay direct the reader to further writings in that topic area, the selection and arrangement of the topics are thoughtful and well-designed. Each author is undoubtedly among the best qualified candidates to undertake his or her commissioned chapter, and most essays are consistently of high quality.

JING ZHU

*Department of Social Sciences
Graduate School of the Chinese Academy of Sciences
Beijing, 100049, China
Email: jingzhu@gscas.ac.cn*

Computational Developmental Psychology

THOMAS R. SHULTZ

Cambridge, MA: MIT Press, 2003

322 pages, ISBN: 026219483x (hbk); \$38.00

Since the publication of Rumelhart and McClelland's *Parallel Distributed Processing* in 1986, critics have debated whether neural network models of learning and development are—at best—video-game-like simulations of real organisms and their nervous systems, or—at worst—clever, hand-built demonstrations that bear only a superficial resemblance to the real-world behaviors that they are intended to represent. In response to this subtle but pervasive skepticism, Thomas Shultz has delivered a comprehensive text that skillfully illustrates the merging of computer models and developmental psychology into a new beast, part of an emerging, interdisciplinary field called “developmental science.”

Shultz's *Computational Developmental Psychology* is organized to achieve three goals. First, it provides an intuitive (but also rigorous) overview of the mathematical

foundations for artificial neural network models. At the beginning of chapter 2, Shultz notes that “I present all the essentials here (one-stop shopping) with sufficient background to enable understanding of key ideas by basically all readers, not just those with an extensive background in neural networks and mathematics” (p. 25; the formal mathematical details and derivations are systematically addressed in the appendix). This is certainly an accurate assessment, but I would also caution that a careful reading of the chapter, while not an absolute requirement, greatly enhances understanding of the models that Shultz presents in later chapters. Skimming, on the other hand, might be comparable to the experience of watching a French movie (without subtitles) after taking a year or two of high school French classes. You get the gist, but just barely.

Chapter 2 spans an impressive range of fundamental topics, including (a) the concept of linear separability, (b) network function in mathematical terms (i.e., propagation of activity between neural units, hereafter called “units”), (c) network architectures and topologies, and (d) learning algorithms. It is at this point that Shultz highlights a major theme of the book: although back-propagation-of-error (hereafter “backprop”) is one of the most common methods for tuning the connection weights of a network (thereby improving its behavior), it also has a number of significant flaws. In particular, backprop nets learn slowly, have static topologies, are vulnerable to catastrophic interference (i.e., new learning wipes out old knowledge), scale up poorly to larger problems, and worst of all, are not biologically plausible.

Fortunately, each of these weaknesses is successfully addressed by an alternative approach, called “cascade-correlation.” Properly speaking, cascade-correlation is a fusion of two distinct innovations, one network-architectural and the other learning-algorithmic. As Shultz carefully explains, the architectural twist is that cascade-correlation nets “grow” or expand, analogous to the processes of synaptogenesis and neurogenesis. As they recruit new internal units, cascade-correlation nets increase their ability to represent important features in their (putative sensory) input. The algorithmic twist, meanwhile, is that Shultz and his colleagues also employ a novel learning algorithm called “quick-prop,” which learns faster than backprop.

Two nice touches are left for the end of chapter 2. First, Shultz offers the reader a brief summary of other related connectionist designs, including simple recurrent, encoder, auto-associator, and feature-mapping networks. Second, in a bit of deliberate irony, Shultz uses a traditional symbolic model to classify 18 different subsymbolic or connectionist models. The result is a decision tree that groups models as a function of common objectives and constraints (e.g., whether the model is *supervised*, i.e., trained with explicit feedback). The maneuver is a little like inviting your atheist friend over for dinner, and then asking them to pray. Shultz’ humor is likely to be appreciated by readers who know the history of “classical” and “modern” AI. In any case, the decision tree (Table 2.2, p. 72) is a valuable aid to novice model-builders.

As a result of the focus on cascade-correlation, there are other notable computational approaches to modeling development that are either excluded or only briefly mentioned by Shultz, including adaptive resonance theory, reinforcement learning, genetic algorithms, and dynamic field theory (e.g., Schlesinger & Parisi, 2004). While none of these are conventional connectionist models, strictly speaking, they still share a common mathematical framework and many theoretical tenets with the connectionist approach (e.g., that knowledge representations are graded and distributed). In chapters 3–5, however, it becomes apparent that Shultz' decision to highlight the virtues of cascade-correlation models is guided by a compelling blend of careful reasoning and broad empirical support. As Shultz seems to implicitly argue, other modeling approaches are no less important or interesting, but simply less successful in capturing three fundamental aspects of development: knowledge representation, developmental transitions, and stages of development.

As Shultz works towards achieving his second goal—to systematically compare the performance of selected models on well-known developmental phenomena—several unique features of the book are revealed. First, the discussion of knowledge representation (i.e., encoding, storage, and retrieval) offers balanced coverage of two prevailing views. On the one hand, the rule-based view characterizes knowledge as symbolic, proposition-like bits of information that are stored in something analogous to a catalogue or lookup table. Connectionism, on the other hand, characterizes knowledge as distributed and graded, rather than symbolic, and represented by the strength of connections between a network of units that are always active to one degree or another.

Shultz surveys a wide array of developmental phenomena in chapter 3, helping the novice model-builder to appreciate both the conceptual and methodological differences that arise from implementing symbolic versus connectionist models. A number of well-known topics are covered, including both logico-mathematical cognition (e.g., number conservation and the balance-scale) and language acquisition (e.g., semantics and morphology). Developmental researchers who specialize in these areas will be more-than-pleasantly surprised to find that Shultz is far from a “bull in the china shop.” Indeed, he goes well beyond a superficial overview of each topic to offer careful, detailed descriptions that are meticulous and thorough. For example, reviews of children's reasoning on Piaget's number conservation and balance-scale tasks include discussion of the (somewhat esoteric) set-size and torque-difference effects, respectively.

The contrast between rule-based and connectionist models also allows Shultz to exploit a second unique feature of the book: the “bakeoff.” As the name suggests, a bakeoff is a head-to-head comparison of two models that are designed to solve the same task. Borrowed from the field of machine learning (e.g., the performance benchmark), the bakeoff is probably an unfamiliar research tool for most experimental psychologists. Nevertheless, there are two crucial lessons that the reader should not miss at this point (Shultz explicitly notes the first; he either implies, or I infer, the second). First, on a practical level, the bakeoff creates a level playing field, on which two alternative accounts can objectively and fairly compete.

Second, on a more symbolic level, the bakeoff also illustrates the need for computational models to be precise and specific in their formulation, in contrast to “traditional” verbal theories, which are more general, and sometimes vague, ambiguous, or difficult to formalize and test.

As Shultz is careful to note, the purpose of the bakeoff is not simply to figure out whether it is the connectionist or rule-based model that learns faster, or better than the other, but rather to perform a much more in-depth *comparative analysis* of the two models. In other words, a comparative analysis highlights the shape of the developmental trajectory, and includes questions such as: Is the process of learning in the model gradual or discontinuous? If discontinuous, are the changes analogous to stages of development in human children? If the model, like children, learns through a series of stages, when do those stage transitions occur, and what learning experiences precede or predict the emergence of a new stage?

Chapter 4, a *tour-de-force* in just under 50 pages, focuses on the issue of developmental transitions, i.e., the “motor” that drives the developmental process forward. To illustrate the topic, Shultz returns to the development of children’s reasoning on the balance-scale and conservation tasks, this time using the bakeoff between symbolic and connectionist models to highlight how and when new behaviors emerge. For example, on the balance-scale task, young children watch as weights are placed on each side of a balance-scale (sometimes at different distances), and then are asked to predict which side (if either) will tip. An intriguing developmental pattern is that young children initially focus exclusively on the magnitude of the weights on each side. With experience, though, they begin to attend to, and incorporate into their reasoning, both the magnitudes and distances of the weights (from the center of the scale). Perhaps not surprisingly, Shultz showcases the virtues of cascade-correlation as the model best-suited for explaining (and sometimes predicting) these changes in behavior and thought.

Shultz not only addresses a number of empirical topics in chapter 4, but several fundamental conceptual issues are also systematically explored. First, both “classical” (i.e., Piagetian) and modern theoretical approaches to developmental transitions are reviewed. Second, Shultz refutes (or at least responds to) “Fodor’s paradox,” part of an analytical, nativist assault on connectionist models which argues that learning and development are a reorganization of prior (read: innate) knowledge, not an acquisition of something new (Fodor & Pylyshyn, 1988). The chapter concludes with two topics that also link psychology and philosophy via epistemology: the relation between learning and development (a distinction developmentalists constantly bicker about), and the relation between evolution and development (i.e., nativism, empiricism, and constructivism). While the latter issue is inevitably a minor theme for Shultz, it should be noted that other authors—in particular, Elman et al. (1996)—have tackled this question in greater detail.

In chapter 5, Shultz combines two lines of attack: the first half of the chapter presents an in-depth analysis of discontinuity in development (i.e., the emergence of qualitatively new behaviors), while the second half emphasizes the now-familiar

blend of behavioral and modeling studies to illustrate the phenomenon of developmental stages and how they are investigated. Shultz' coverage of *functional data analysis*—while arguably the appropriate tool for identifying qualitative changes—runs a bit jargon-heavy (e.g., velocity peaks, B-spline basis function, curvature of growth, etc.), and may only stir the souls of true math-lovers. In this case, the topic is an admittedly dry one, and so there may be no sure remedy for grabbing the attention of the more casual reader. However, the latter half of chapter 5 allows Shultz to focus on a number of concrete examples, including development on both traditional cognitive tasks (e.g., conservation, balance-scale, and seriation) and linguistic measures (e.g., pronoun use and phoneme discrimination). Shultz succeeds in finishing the chapter on a high note for developmental psychologists, as the question of developmental precursors (i.e., how prior structures influence the emergence of later ones) is interpreted and addressed in computational terms.

The final section of the book (ch. 6–7) represents the culmination of Shultz's third and most laudable goal: first, to raise and then respond to the wide array of issues that critics have raised against computational modeling in general and connectionism in particular, and second, to forecast where the field of computational developmental psychology is headed. The manner in which Shultz dispatches the critics of connectionism is especially fun to watch, as he incorporates a compelling mix of empirical data, logical analysis, and just plain common sense to rebut the skeptics. Meanwhile, the last chapter highlights numerous emerging trends, each of which is likely more than a passing fad, and will inevitably exert some influence on how computational models of development are designed and studied in the next 20 years.

References

- 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.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.
- Schlesinger, M., & Parisi, D. (Eds.). (2004). Beyond backprop: Emerging trends in connectionist models of development. *Developmental Science*, 7, 131–132.

MATTHEW SCHLESINGER
Psychology Department
Southern Illinois University
Carbondale, IL 62901, USA
Email: matthews@siu.edu

Three Faces of Desire

TIMOTHY SCHROEDER

Oxford, England: Oxford University Press, 2004

224 pages, ISBN: 019517237X (hbk); \$35.00

Three Faces of Desire is a refreshing reexamination of desire theory, guided by evidence from behavioral and cognitive neuroscience. Where prior theories have taken for granted that either motivation or pleasure is the essence of desire, this book construes pleasure and motivation as effects of desire. Against motivational and hedonic views, Schroeder argues for a “Reward Theory of Desire”: desire is the capacity to represent a state of affairs such that the representation contributes to reward-driven learning (p. 131). Aversion is given a similar treatment with punishment in the place of reward.

A substantial portion of the book argues against leading accounts of desire. Schroeder groups the opposing views under two headings: the standard account, and the hedonic account. According to the standard account, desire is a disposition to bring about a state of affairs. Under this view, possessing the desire that I eat a sandwich entails that I will be disposed to make it true that I eat a sandwich. According to the hedonic view, desires are dispositions to feel pleasure or displeasure when a state of affairs seems to hold. Under this view, the desire that I eat noodles disposes me to take pleasure in eating noodles. Schroeder considers numerous variations on the standard and hedonic theories and summons philosophical and empirical arguments against both. His treatment of both views is very thoughtful, and his objections are numerous and carefully stated.

According to Schroeder, common sense holds that we can have desires that do not motivate us to make their contents true, such as the desire that it be sunny tomorrow. That is not to say that such a desire will have no motivational implications, but rather that the content of the desire does not specify the action. Those of us who realize that we have no power over the weather can continue to desire sunny days, and because of such desires, we can be motivated to tune into the local weather report. But the desire that the sky be sunny does not have the constitutive function of making us try to bring forth the sun, or to do anything for that matter. So, desires are not essentially motivational states.

Among his objections to the hedonic theory, Schroeder offers an argument on the basis of deficit evidence (pp. 33–34). People who have had a portion of their anterior cingulate removed have the ability to desire intact (i.e., they exhibit preferences), but their ability to feel pleasure and displeasure is severely impaired. So, desire can be dissociated from both motivation and pleasure. Although desires can cause us to act, they are not essentially motivational states. Though they may be required for us to feel pleasure, pleasure is not their essential feature either.

The third face of desire—and its essence—is its role in determining what will count as a reward. Chapter 2 clarifies the sense of reward Schroeder has in mind, as well as the appropriate sense of punishment for understanding aversion. The old behaviorist doctrine held that a reward is a stimulus that, following a behavior,

increases the likelihood that an animal will behave in the same way in similar circumstances. Schroeder rejects this view, which he takes to be an expression of the standard theory of desire. We should not identify reward and punishment with pleasure or displeasure either, because pleasure and displeasure result from reward and punishment. Rather, a reward is an event or thing whose representation tends to produce a certain kind of neural network training signal. This training signal is described by a class of learning algorithms that fall under the heading of reinforcement learning—which is not to be confused with some form of behaviorism (Sutton & Barto, 1995).

Chapter 3 argues that pleasure and displeasure are representations that track desire satisfaction. Pleasure is felt when an agent represents a net increase in desire satisfaction greater than what's expected (p. 90). A similar account follows for displeasure. Alternative accounts of pleasure view it as a modification of behavioral dispositions, clusters of bodily sensation, a distinctive kind of quale, or a player of particular functional roles. Schroeder measures his view against the alternatives by appeal to a set of facts that a theory of pleasure should explain. These include the abnormality of experiencing both pleasure and displeasure simultaneously, the graded nature of hedonic tone, and the manner in which pleasure appears to drive out displeasure. Schroeder explains the first fact by appeal to a dual systems view of reward and punishment. If separate systems realize reward and punishment (in this case separate dopamine and serotonin pathways) then simultaneous sensations of pleasure and displeasure could be generated by opponent processes. The second and third facts can be explained by pleasure and displeasure having contrary content, like negative and positive numbers, ranging over a spectrum of strengths, where at the neutral point desire satisfaction is represented as neither increasing nor decreasing. One should note that Schroeder's theory of pleasure has the counter-intuitive consequence that pleasure should only be experienced when one is in some sense surprised.

In Chapter 5, Schroeder elaborates the reward theory of desire (RTD), attending primarily to the kind of learning that desires are supposed to produce. The key insight comes from the neuroscience of midbrain dopamine signaling, largely generated by Wolfram Schultz's laboratory. Through Schultz's work (e.g., 2002), a number of computational neuroscientists have come to believe that mesolimbic and nigrostriatal dopamine cells produce an analogue to the system training signal that reinforcement learning theory calls "error in predicted reward." Essentially this signal reports the difference between a prediction of reward value and the reward actually received, which in the reinforcement-learning literature typically depends on an action taken by a learner. When the difference between representations of actual and predicted reward is nonzero, weights in a reinforcement-learning network are adjusted. Schroeder proposes that aversions are distinct from desires. A desire with positive valence predicts that a represented state of affairs will eventuate in positive reward return. An aversion predicts that a represented state of affairs will subtract from current reward. So, in Schroeder's hands, reinforcement learning systems work

to optimize total reward from both directions—minimizing punishment and maximizing reward, presumably, over a variety of timescales.

RTD, unlike typical reinforcement learning models, does not require desires to drive action. Changes in perceptual discrimination, attention, declarative memory and other cognitive functions can all be driven by reinforcement learning. This is an interesting claim, since dopamine pathways appear to mediate the additional forms of plasticity that Schroeder describes.

In Chapter 4, Schroeder fleshes out an explanation of how desire contributes to goal-directed behavior without being essentially constituted by that function. Schroeder accepts that desire can be a source of motivation, but it is not the only source. Prior intentions, trying, pleasure, and awareness of reward can all be sources of motivation. Interestingly, pleasure, in Schroeder's view, is much less a contributor to movement than common sense would suggest.

Not surprisingly, desire influences movement by determining what will count as a reward. The picture is as follows: sensory representations lead to the formation of motor programs, or patterns of neural activity that are poised to cause specific movements. After they are formed, these motor programs are held in inhibition. The same sensory representations that triggered the formation of motor programs are simultaneously decoded to determine what kind of rewards the environment contains, relative to expectations. Of those motor programs held in suspension, the ones associated with the highest reward value are unleashed and goal-directed movement commences.

So, the three faces of desire—motivation, pleasure, and reward—are all accounted for, and though motivation is not necessarily dependent upon desire, pleasure is. RTD is not as easy a theory to grasp as the standard and hedonic theories, though it appears to be the most biologically plausible. Of course, the standard and hedonic theories do not exhaust the alternatives to RTD, nor are they the most recent accounts of desire to come out of the literature. But work on the pro-attitudes has been at the periphery of interest among philosophers of mind for some time now, making Schroeder's critique of the standard and hedonic theories a valuable review.

Though the book is well organized and much of it is clearly written, sometimes the connections between Schroeder's theory of desire and the formal theory that inspires it are less easy to follow. For example, there are no sample algorithms to give a sense of how to understand reinforcement formally. Schroeder only mentions that there are different kinds of algorithms out there, and though this is true, his reliance on the formal theory might have warranted an appendix. Reinforcement learning theory does a lot of work in *Three Faces*, but little attention is directed toward explaining what it is. This is likely the result of the book's focus on the neurobiological results that have been interpreted in light of reinforcement learning.

Readers familiar with the computational aspects of reinforcement learning may be confused by the nonstandard use of terminology. Schroeder makes no mention of the basic components of reinforcement learning systems: the reward function, value function, and the policy (Montague, Eagleman, McClure, & Berns, 2002). Briefly,

a reward function assigns real number values to all states of the system (e.g., the value of stepping on an electrically charged plate, which results in a shock). A value function assigns values to all states of the system given a projected succession of states (e.g., the value of stepping on an electrically charged plate, given that food is on the other side). A policy maps sensory states to actions. The goal of a reinforcement learning system is to maximize total reward.

The absence of any discussion of these fundamentals makes it difficult to understand how reinforcement learning is supposed to influence the spectrum of functions that Schroeder associates with desire and dispenses with a useful tool for distinguishing various aspects of desire when construed as a component in reinforcement learning systems. Other peculiarities result from under-describing the formal theory. For instance, two very different kinds of phenomena are called “reward” in the book. There are the events whose frequency and magnitude are tracked by desire (e.g., food rewards). Then, there are reinforcement signals that result from errors in predicting returns for those tracked events. Only the former is regarded as a reward signal in reinforcement learning theory. The author does not appear to be aware of this issue, but does not introduce notation or new terms to help keep the distinction clear.

The neuroscience in the book is just enough to get Schroeder’s ideas across. Those interested in more neurophysiology might look to Rolls (2005) to supplement their reading. But Schroeder is not stingy. By the end of the book, Schroeder invokes a variety of interactions in the brain to explain how desire fits with pleasure and motivation. Discussions of function include midbrain dopamine in reward (the ventral tegmental area and pars compacta of the substantia nigra), serotonin in punishment (the dorsal raphe nucleus), the perigenual region of the anterior cingulate in the processing of pleasure and displeasure, the supplementary motor area and the motor region of the anterior cingulate in trying, the primary motor cortex in basic actions, and prefrontal cortex in the formation of short-term prior intentions. The organization of these claims could have been greatly enhanced by the inclusion of box-and-arrow diagrams, though even readers with little exposure to neuroscience will probably be able to sketch such diagrams as they work through the text.

Three Faces of Desire is a work of neurophilosophy that does not short-change the analytic tradition. There are a lot of details to be filled in, but Schroeder’s arguments are always lucid and often compelling. *Three Faces of Desire* introduces a rich body of neuroscientific data that has been neglected by philosophy of mind for some time, but because of the use Schroeder has made of it, it is doubtful that this neglect will persist for long.

References

- Montague, P. R., Eagleman, D. M., McClure, S. M., & Berns, G. S. (2002). Reinforcement learning. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 908–913). London: MacMillan.

Rolls, E. T. (2005). *Emotion explained*. Oxford, England: Oxford University Press.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36, 241–263.

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

ANTHONY LANDRETH
Department of Philosophy
University of Cincinnati
Cincinnati, OH 45221-0374, USA
Email: phi@infinitenavel.com