PAUL M. CHURCHLAND

Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals

Paul M. Churchland, *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*, MIT Press, 2012, 299pp., $35.00 (hbk), ISBN 9780262016865.

Reviewed by Peter Carruthers and J. Brendan Ritchie, University of Maryland

Plato's Camera is Churchland's most recent presentation of his neurally-inspired theories of learning and mental representation, together with their bearing on issues in epistemology and the philosophy of science. The book is clearly and engagingly written, and revisits many of the debates that Churchland has engaged in over the course of the last 30 years or so.

Following an introductory chapter outlining his position, the next two quite lengthy chapters are intended to provide a neurally realistic theory of what Churchland calls "first-level learning." This is the sort of learning that issues in knowledge of the structural and causal invariances of the world (which inspires the book's title), and which results from the gradual alteration of synaptic weights between neurons. While this idea is first explained in terms of error-correcting algorithms (discussed in Chapter 2), it is later replaced with the more neutrally-plausible mechanism of Hebbian plasticity ("neurons that fire together wire together"; discussed in Chapter 3). Either way, the result is said to be a large set of neural populations, each of which has been sculpted into a high-dimensional feature-map of some domain. (In fact, Churchland repeatedly returns to the idea of the mind as comprised of multiple maps.) In the case of face-recognition, for example, the neural population in question will represent the various dimensions along which faces can differ from one another. Recognition of a particular face will then result from heightened activity in a specific region of this state-space, which lies at the intersection of activation-levels along each of the dimensions of the space that are reliably evoked by the face in question.

Chapter 4 then deals with what Churchland calls “second-level learning”, which involves the redeployment of existing conceptual frameworks to new domains. Here Churchland provides an account of paradigm-changing forms of scientific discovery. For example, he dwells at length on Newton’s insight that the force that causes apples to fall to the ground might explain the orbit of the moon around the Earth (p.192-194). Discussing a number of other famous examples, Churchland builds toward a vindication of scientific realism, intended to rebut famous anti-realist objections such as the pessimistic meta-induction and the underdetermination of theory by evidence. While the second objection receives a very lengthy discussion that we cannot examine here, the rough idea of Churchland’s reply to the former is that the history of science consists of us replacing partially correct maps with more explanatory, more predictive maps. He thus proposes an *optimistic* meta-induction, to the effect that we can expect future theories to consist maps that are ever more accurate than those partially-correct ones we use today.

Chapter 5 describes third-level learning, which is said to be uniquely human. This involves cultural learning (especially using the resources of natural language) as well as collective cognitive activity involving communication and debate. One thing Churchland emphasizes is how much this sort of third-level learning has involved the invention of regulatory mechanisms to improve learning at the second level and to allow for more effective transmission of knowledge. These include everything from record-keeping to shared standards of epistemic evaluation.

That, in broad outline, is the framework defended in the book. What does Churchland say about opposing positions on these topics? While the book has many virtues, it is unfortunate that Churchland repeatedly fails to do justice to his opponents’ views. For the most part he critiques implausible caricatures, rather than engaging sympathetically but critically with the most charitable interpretations of their positions. Three examples (to be discussed in more detail in the following paragraphs) are his criticisms of nativists, his repudiation of the language of thought hypothesis, or “LOT” (championed by Fodor and others), and his critique of indicator semantics (of the sort defended by Dretske, Fodor, and others). Moreover, although Churchland has many insightful things to say in the final two chapters of the book about the question of scientific realism and the role of natural language and cultural institutions in shaping some aspects of human cognition, it remains far from clear that adherence to anything resembling his neurally-

inspired story is needed for one to say those things. The result is a deeply unsatisfying book. Churchland has missed an opportunity to show us, not only that his neural state-space account is actually inconsistent with nativism and LOT theories, but also that it has genuine advantages that opposing views cannot accommodate.

The state-space theory of the most basic level of representation in the brain is by no means implausible. Indeed, the idea of distributed-representation neural networks is very popular in cognitive science. But there is nothing in such an account itself to exclude a significant role for innateness. Churchland is resolute in opposing any such role, however, citing the small number of genes contained in the human genome when compared with the astronomical number of neural connections. But no actual nativist thinks that individual neural connections are directly coded for in the genome. Rather, all believe that innate systems result from interactions between genes, developmental variables, and environmental influences. To a first approximation, the minimal commitment of nativists in cognitive science is that some features of our neural and cognitive systems are acquired or develop *without learning*, rather than that they are directly coded in the genes (Carruthers et al., 2005, 2006, 2007).

Churchland makes no attempt to engage with the actual views of real nativists, nor with the sorts of empirical data that motivate their views. For example, we now know that face processing in both humans and macaque monkeys is undertaken in an intricately interconnected set of six cortical regions, which appear to be homologous across the two species (Moeller et al., 2008; Tsao et al., 2008). We also know that both human and monkey infants have the capacity to distinguish between faces and non-faces (such as scrambled facial components) at birth (Farroni et al., 2005). Moreover, monkeys who have never had any exposure to faces at all (who were raised by humans wearing opaque gauze masks over their heads) nevertheless show capacities for fine-grained discrimination among both human and monkey faces that are close to normal (Sugita, 2008).

What such data suggest is that primates possess an innately channeled domain-specific learning mechanism specialized for faces, which can perform at least some aspects of its function without learning. Moreover, there is extensive evidence supporting the existence of many such mechanisms in humans and other animals. Many animals can walk from birth, for example, and

are already capable of representing a good deal about the spatial and causal structure of the world around them. That the same *seems* not to be true of human infants may result more from the highly altricial nature of human infancy—since the heads of human infants would otherwise be too large to travel down the birth canal—rather than from the absence of innate learning mechanisms. Indeed, views of this sort are defended by those who have used looking-time methods to reveal the existence of a number of different bodies of so-called "core knowledge" in human infants (Spelke and Kinzler, 2007).

In addition, Churchland makes no mention of the many instances of one-shot learning that are known to exist in the animal kingdom, although he himself emphasizes the slow pace of connectionist and Hebbian learning. For example, a bee needs to observe the dance of compatriot just once to know the direction and distance of a nectar source, and a baboon can come to know the new rank-ordering of families and individuals in the troupe from overhearing a single agonistic exchange that concludes with a rank-reversing fear scream (Cheney and Seyfarth, 2007). It may be that such findings can be explained in connectionist or Hebbian terms, but Churchland does not attempt to tell us how.

Failure to engage with his actual opponents is equally characteristic of Churchland's discussion of LOT theories. He writes disparagingly of such accounts: "Encouraged further by the structure of our own dearly beloved Folk Psychology, [supporters of LOT] have wrongly read back *into* the objective phenomenon of cognition-in-general a historically *accidental* structure that is idiosyncratic to a single species of animal (namely, humans), and which is of profoundly secondary importance even there" (p.5). Such a claim deeply misunderstands the LOT hypothesis, however. For an appeal to folk psychology is entirely inessential to the motivation for LOT theories, and the claim that such theories try to understand the representational structure of the mind by analogy to human public language is patently false in the case of Fodor (who is, of course, the archetypal LOT theorist).

Furthermore, LOT theories do *not* claim that "sentences" in the language of thought are: "just hidden, inward versions of the *linguistic* representations and activities so characteristic of cognitive activity at the third level [the level of explicit reasoning and communication in natural language sentences, discussed in Chapter 5]" (p.26), a view that Churchland attributes to Fodor

(1975). On the contrary, LOT representations are held by Fodor to be language-*like* only in the sense that they have a combinatorial syntax and semantics and meet the conditions of systematicity and compositionality (Fodor and Pylyshyn, 1988). Mental representations, on a LOT account, are built up out of representational components in such a way that these components make systematic contributions to the representational properties of the complexes in which they are embedded. It is of course true that human language is compositional and systematic. But LOT is not the claim that we have an internal representational system that is merely an internal version of an external, public, language. Nor does anyone believe that LOT is distinctively human, as Churchland claims. On the contrary, many of the kinds of data that are thought to support it derive from the study of nonhuman animals (Gallistel, 1990; Gallistel and King, 2009).

Ironically, Churchland's own account needs to be heavily supplemented to explain the full range of human and animal cognition, and the most obvious supplementation available would introduce LOT representations (properly understood, as above) into the story. Churchland contrasts the conceptual frameworks that result slowly from learning, and reflect the fixed causal structure of the environment, with ephemeral activations within those networks that locate the organism in the here and now, enabling it to know what to expect next or how to effect changes in that environment. But there is a huge space of forms of representation of the environment that is missing from this dichotomy, including both semantic and episodic forms of memory. It is surprising that Churchland could write the entire book without discussing any such examples.

The state-space structures that are thought to be built slowly by Hebbian learning correspond most closely to what would normally be described as *implicit* forms of knowledge. Our knowledge of the ways in which faces vary from one another is mostly implicit and inarticulable, for example. (Indeed, nativists could plausibly appropriate the state-space idea to characterize the internal processing structures of the learning mechanisms that they postulate.) This is the "landscape of abstract universals" described by the book's subtitle. And then online activity of specific regions in these state-spaces represent the here-and-now, such as the face of a specific individual person whom one is now seeing. Yet humans and other animals possess many forms of knowledge that fall into neither of these categories, since they require interactions *between* neural maps. Moreover, these are forms of knowledge that cannot be assimilated to learning at

Churchland's second level (roughly, reasoning by analogy) nor at the third (where natural-language sentences play an important role).

Consider episodic memory, for example. Such memories are not regions in any one state-space. Rather, they seem to involve the creation of long-term linkages between regions of many different state-spaces, corresponding to the various sensory components of the original experience, in such a way that activations of any one are likely to cause activations of the others. If one recalls an episode of three red tomatoes falling on one's kitchen floor and smashing, for example, then this would seem to require a long-term link between the region of color state-space that represents red and the region of fruit-and-vegetable-space that represents tomatoes, together with the region representing a numerosity of three and the region of location-space that corresponds to one's kitchen. Indeed, it is in just such terms that the formation of episodic memory is characterized by many cognitive scientists (Tulving, 2002). But notice that the resulting structure is discrete and distinct from most other episodic memories. It is also compositionally structured out of the state-space regions that represent the various components of the original event.

Something similar will surely be true of many forms of semantic (or "factual") memory. Consider what takes place when one happens to run into a colleague while out walking the dog, and she points out the house where she lives nearby. The resulting knowledge is not comfortably assimilated to knowledge of the enduring causal structure of the world represented by state-spaces themselves. (Nor is the knowledge analogical in nature or natural-language-based.) Rather, it would seem to require building a link from the regions of various state-spaces (e.g. of the face-recognition system) that represent one's colleague to the region of spatial state-space that corresponds to the location of her home. And this, too, will be a compositionally structured discrete representation: a sentence in the language of thought, no less!

One place where there might seem to be a clear contrast with opposing views is on the topic of representational content. Here Churchland defends an updated version of his state-space semantics and contrasts it with "indicator" views such as the well-known positions of Dretske (1988) and Fodor (1990). But again Churchland deals less than sympathetically with his opponents. For example, he objects against Fodor that there are no laws of nature linking such

worldly items as socks with any given state of the brain (p.95). But this is probably to take Fodor's words more strictly than intended. All Fodor need really be committed to is the existence of a reliable causal connection between the two that satisfies his famous "asymmetric dependence" requirement. For he is quite explicit that many other causal processes, many of which might involve representations, can factor into the causal processes that determine a symbol's content (Fodor, 1990, p.110). The crucial point is that the *contents* of these other symbols do not contribute to the content of the symbol in question.

In contrasting state-space semantics with indicator semantics Churchland advances the principle, "No representation without at least some comprehension" (p.96). Here the contrast with Fodor's views is fair, since the latter has always defended a resolutely atomistic account of content. But it overlooks the fact that many theorists who endorse some or other version of indicator semantics think that it forms just one factor in a two-factor account of semantic content, the other factor being some form of what Millikan (1984) calls "consumer semantics", such as her own teleosemantics or a version of inferential-role semantics (Block, 1986). And such theorists, for all Churchland has said, might happily adopt his account of state-space semantics as providing a story about the vehicles of content, and also an account of how the relevant form of teleological or inferential role is fixed.

A final striking fact about Churchland's book is that it seems almost wholly divorced from empirical psychology. Remarkably, indeed, in a book that advances a theory of the mind that is supposed to be empirically supported, Churchland provides only around thirty scientific references, just a third of which date from the twenty-first century, and many of which are computational rather than experimental in nature. One would like to think that he chose to provide only a judicious selection so as not to overwhelm his audience with references. But since he ignores a great many results that appear inconsistent with his main theses, we fear that the paucity of references requires a different explanation. Indeed, Churchland ignores almost entirely the extensive work in developmental and experimental psychology, in neuroscience, and in studies of comparative cognition that have been conducted by cognitive scientists, especially over the last twenty years. And it is precisely once we examine the theories supported by empirical phenomena of these psychological sorts that past and present arguments for nativism and for LOT (appropriately understood) begin to emerge.

References

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10, 615-678.

Carruthers, P., Laurence, S., and Stich, S., eds. (2005, 2006, 2007). *The Innate Mind, Vols. 1, 2, and 3*. Oxford University Press.

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

Dretske, F. (1988). *Explaining Behavior*. MIT Press.

Farroni, T., Johnson, M., Menon, E., Zulian, L., Faraguna, D., and Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effect of contrast polarity. *Proceedings of the National Academy of Sciences*, 102, 17245-17250.

Fodor, J. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture. *Cognition*, 28, 3-71.

Gallistel, R. (1990). *The Organization of Learning*. MIT Press.

Gallistel, R. and King, A. (2009). *Memory and the Computational Brain*. Blackwell.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. MIT Press.

Moeller, S., Friewald, W., and Tsao, D. (2008). Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science*, 320, 1355-1359.

Spelke, E. and Kinzler, K. (2007). Core knowledge. *Developmental Science*, 10, 89-96.

Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences*, 105, 394-398.

Tsao, D., Moeller, S., and Freiwald, W. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105, 19514-19519.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, 53, 1-25.