Review of Paul M. Churchland, Plato's Camera

Citation for published version:
Isaac, AMC 2014, 'Review of Paul M. Churchland, Plato's Camera' Philosophy of Science, vol 81, no. 1, pp. 161-165., 10.1086/674366

Digital Object Identifier (DOI):
10.1086/674366

Published In:
Philosophy of Science

Book Review

Paul M. Churchland, ***Plato's Camera***. Cambridge, MA: MIT Press (2012).

Alistair M. C. Isaac
University of Edinburgh

Analytic philosophy has been dominated by the idea that analogies with language provide the appropriate starting point for analyzing the mind, the world, and the relationship between the two. In *Plato's Camera*, Paul Churchland provides a compelling corrective to this trend, attempting to counteract its "cognitive inertia" (276) by providing a coherent and thoroughgoing alternative framework, one rooted in an analysis of mental representations as picture-like rather than language-like. This framework synthesizes a number of strands which wind through his entire career, including connectionism, realism, semantic holism, and the plasticity of mind. As such, it constitutes a definitive statement of Churchland's mature views across epistemology, philosophy of mind, and philosophy of science.

Churchland's account rests on his understanding of the representational capacity of artificial neural networks (ANNs). An ANN is a collection of simple units, or "nodes," linked together by edges weighted with real values. Information is input to the network by activating some subset of the nodes; the signal from these nodes then propagates through the network, attenuated by the weights on each edge. Just like biological neurons, each node integrates over the inputs which arrive at it through connecting edges, then "fires," i.e. itself sends out a signal of some real value depending upon the total value of this weighted input. ANNs are typically initialized with random edge weights, then weights are adjusted to achieve the desired output on a categorization task during a period of training on a labeled data set. Given the apparent similarities between an ANN and the neural structure of the brain, the question of whether ANNs represent information in a qualitatively different manner from (language-like) symbolic systems has been an important one in philosophy of cognitive science, a question which Churchland answers convincingly in the affirmative in *Plato's Camera*.

More specifically, the *n* nodes in the middle, or "hidden," layer of a three-layer ANN define an *n*-dimensional activation space, with each dimension corresponding to the degree of activation of a single node. Any input to the network induces a vector in this activation space, and analysis of trained networks reveals that categories in the input correspond to regions in this activation space. This is true even when the categories have not been explicitly given to the network during training, i.e. the data set is not labeled, as in the case of "unsupervised learning." Using this technique, connection weights are adjusted until the network reproduces the input vectors induced by a set of exemplars at its output layer. If there are fewer hidden nodes than input and output nodes, the process of compressing information in the input into the hidden layer extracts patterns in the training set, i.e. the network "discovers" for itself categories in the data.

For instance, in one of Churchland's favorite examples, post-training analysis of a network trained to reproduce pixelated images of faces revealed that the hidden layer had separated male and female faces into distinct regions in its activation space, as well as grouped faces of the same individual together (Cottrell, 1991).

Churchland takes this analysis of artificial neural networks to reveal how the world is successfully represented in realistic brains (both human and animal). High-dimensional neural activation spaces provide a "picture" or "map" of the structure in the world which induced them. These spaces are shaped during "first-level learning," the gradual alteration of synaptic weights in response to environmental feedback, generating a hierarchical conceptual framework, where "concepts" are understood as basins of attraction within the neural activation space. It is this hierarchy of concepts which provides a map of "the timeless and invariant *background structure* of the . . . physical universe" (viii). This is the sense in which the brain is "Plato's camera": it veridically images the true categorical structure of the world before it. This neurally-based account of conceptual structure applies equally to animal cognition, demonstrating an evolutionary continuity problematic for typical linguacentric analyses of concepts.

After a short survey chapter, most of the first two-thirds of the book comprises a careful elucidation of first-level learning and its philosophical consequences in two long chapters. It is in these chapters that Churchland develops and defends his theory of abduction, his holistic "domain-portrayal semantics," and his realism, as well as contrasting his views with classical positions in the literature, particularly those of Peirce and Kant. Here also is Churchland's most direct engagement with a contemporary philosophical opponent, namely Jerry Fodor, whose linguacentric and anti-holistic views provide a convenient foil for domain-portrayal semantics. Pace Fodor, Churchland clearly and convincingly articulates why there are no atoms of meaning in neural network representations, thereby clearing up a point of much confusion in the debate between connectionists and advocates of symbolic systems (84–90).

Next comes a chapter on "second-level learning," or the process by which existing conceptual maps are applied to new topics. We understand the state of the world by locating ourselves within a conceptual map, "indexing" it. This location is determined in part by occurrent sensory input, but is also modulated via recurrent connections by the current dynamical state of the brain. Through this self-modulation, we are able to make analogical leaps, applying pre-existing conceptual maps in novel ways to different problems. Churchland uses this idea to analyze examples from the history of science, developing his account of scientific theories and theory change. On his view, theories are neurally-defined conceptual frameworks, typically so complicated that the scientist herself does not have explicit enough access to express the full details of her expertise sententially (hence Churchland's break with syntactic and semantic accounts). It is not the theory itself, but the methodology by which it is assessed, which defines such a conceptual framework as scientific, namely evaluation in terms of "overall success in repeatedly generating local expectations that agree with the subsequent spontaneous or automatic indexings of its high-dimensional categories" (202). The mark of scientific frameworks against other (e.g. religious) conceptual frameworks can be seen in their patterns of convergence and unification over time (234).

The shortest, concluding chapter covers "third-level learning," or cultural change via language. While language is not fundamental to cognition, it nevertheless has the power to regulate cognitive activity. The results of first- and second-level learning can be preserved and transmitted via language, allowing learning on a cultural level—human society as a whole increases its knowledge of the world, storing, regulating, and transmitting that knowledge via an expanding network of cultural institutions. Churchland acknowledges here resonances of Hegel, society itself as a growing, evolving, and metabolizing entity, though he rejects the specific analogy between society and our pre-theoretic understanding of mind as an insufficiently rich basis for detailed theorizing (261, 277).

Throughout the book, Churchland's argument relies heavily on suggestive analogies and diagrams. This stylistic choice is a double-edged sword. On the one hand, it creates a feeling of quick intuitive understanding in the reader. On the other, however, I worry that at times it obscures problem areas for Churchland's view. One example is his analysis of identity and similarity relations between high-dimensional spaces (104–115). A theory of such relations is necessary for two critical steps in Churchland's argument: first, to counter the worry that holists cannot account for sameness of meaning across different individuals; second, to define the fit between neurally-defined conceptual structure and external "objective" structure in the world on which Churchland's realism depends. Yet Churchland's discussion of this critical problem is very flip (his words: "agreeably simple" (112)) and rests heavily on our intuitions about its solution for the case of two-dimensional maps (i.e. via superposition and rotation). Even in the two-dimensional case, however, there are well-known problems for the "partial homomorphism" view Churchland offers—e.g. some, but not all, hexagons represent France. Furthermore, Churchland's account does not obviously generalize to the comparison of structures with different dimensionality, yet this is what would be required to assess sameness of meaning across individuals, since typically the exact number of neurons involved in a representation will differ for different brains.

Even if Churchland's analysis of similarity of structure succeeds, it is unclear that it is sufficient to motivate his realism. Churchland argues that the "slippery slope to Idealism" begins with the erroneous view that mental representation rests on first-order resemblance (79, 82). Churchland in contrast embraces second-order, or structural, resemblance as the foundation of mental (and scientific) representation (81). But does the turn toward second-order resemblance really justify his frequent references to external structure as "timeless and invariant" (viii), "objective" (76*f*), and "enduring" (215)? It seems the more natural interpretation is the pragmatic one which also appears at various points in Churchland's rhetoric, as when he rejects a sharp distinction between "factual knowledge" and "practical skills" (32) or advocates evaluating theories on the basis of "internal consistency" and "overall success" (234). It would be entirely consistent with Churchland's neural arguments to embrace this pragmatism, motivating an epistemology far closer to that of Peirce than he admits.

This raises a final concern about the book, namely its relative lack of engagement with recent literature, both philosophical and scientific. This is perhaps the prerogative of a distinguished philosopher with a systematic view, yet at times a more explicit engagement with related views would have helped to clarify and legitimate Churchland's position. For instance, much of what he says on the topic of scientific realism resonates with the recent claims of so-called structural realists. No doubt Churchland would reject their reliance on the semantic view of theories, but would he endorse their analysis of the theory-world relation? An explicit discussion would have been useful. In terms of relevant science, Churchland relies heavily on research in artificial neural networks dating to the late '80s and early '90s, but connects little with more recent developments. To mention one significant omission, Bayesianism has become prominent recently as a paradigm both in psychology (e.g. Gopnik, 2012) and in neuroscience (e.g. Ma et al., 2006), to say nothing of its popularity in philosophy of science, yet it receives no discussion here. These omissions do not undermine the value of Churchland's views, but they may constitute a stumbling block for the impatient reader.

Plato's Camera has much to recommend it. As the mature position statement of one of the most significant and influential philosophers of cognitive science, it is required reading for both cognitive scientists and philosophers of cognitive science. Furthermore, it will be of value to philosophers of science, epistemologists, and philosophers of mind receptive to the idea that neuroscience may somehow inform their respective fields. The discussions of abductive inference (68–73), mental representation (74–104), and the underdetermination of theory by evidence (215–223) are especially valuable.

Bibliography

Cottrell, Garrison W. (1991) "Extracting Features from Faces using Compression Networks: Face, Identity, Emotion, and Gender Recognition using Holons," in *Connectionist Models: Proceedings of the 1990 Summer School*, ed. D. Touretsky et al. San Mateo: Morgan Kaufmann (328–337).

Gopnik, Alison (2012) "Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications," in *Science* 28: 1623–1627.

Ma, Wei Ji, Beck, Jeffrey M., Latham, Peter E., and Pouget, Alexandre (2006) "Bayesian Inference with Probabilistic Population Codes," in *Nature Neuroscience* 9: 1432–1438.