Topoi (2006) 25: 29–32
DOI 10.1007/s11245-006-0024-z

Into the brain: where philosophy should go from here

Paul M. Churchland

Abstract The maturation of the cognitive neurosciences will throw light on many central philosophical issues. Among them: semantic theory, perception, learning, social and moral knowledge, and practical reasoning and decision making. As contemporary medicine cannot do without the achievements of modern biology, philosophy would be pitiful if it disregarded the achievements of brain research.

Keywords cognitive neurosciences · semantic theory · perception · moral knowledge · practical reasoning · Hebbian learning

The physical brain, of both humans and animals, has begun to give up its secrets. Those secrets have been locked away in a bony vault, encrypted in a microscopic matrix of 100 billion neurons and 100 trillion synaptic connections, for the entire history of our philosophical musings, with no more influence, on the content of those musings, than the influence exerted by the equally hidden secrets of the kidney, or of the pancreas. The winding path of our philosophical theorizing has been steered by other factors entirely. Those factors have been many and various, even glorious, and they have been precious for existing at all. But they have not included even the feeblest conception of how the biological brain embodies information about the world, or of how it processes that information so as to steer its biological body through a complex physical and social environment. In these dimensions, we have been flying blind for at least three millennia.

P. M. Churchland (✉)
University of California, San Diego, CA, USA
e-mail: pchurchland@ucsd.edu

But our blinders here have begun to be lifted, and our ignorance has begun to recede. A new generation of techniques and machines of observation has given us eyes to see into the encrypted details of neuronal activity. A new generation of scientists has given us a self-critical community of determined empirical researchers. And a new generation of theories has given us at least an opening grip on how the brain's massive but microscopic matrix might perform the breathtaking feats of real-time cognition that so compel our philosophical attention. My aim in this short paper is to outline the various ways in which the maturation of the cognitive neurosciences is likely to throw light on an unprecedented variety of issues of central and historical importance to philosophers in particular, issues near and dear to all of us, issues that have long defined our field. The overall impact of the maturing neurosciences, most will politely allow, is likely to be substantial. But most philosophers, I'll wager, expect the impact on philosophy to be relatively minor, if they have any expectations on the matter at all. How mistaken they are is the topic of this short paper.

Let me begin in what may be an unexpected place: semantic theory. How does the brain *represent* the enduring structure of the world in which it lives? The emerging answer, it seems, is surprisingly Platonic. The brain slowly develops, by a process to be discussed below, a high-dimensional *map* of the abstract categories, invariant profiles, and enduring symmetries that provide the unchanging *background structure* of the world of ephemeral processes. More accurately, the brain develops a substantial number of such maps, each one of which represents a specific domain of contrasting but interrelated universals, such as the domain of colors, the domain of voices, the domain of shapes, the

domain of motions, the domain of animals, and so forth. Each map contains an appropriate family of *prototype* positions for each family of learned categories, and the assembled *proximity and distance relations* that configure those prototype positions within the map are collectively homomorphic with the assembled *similarity and difference relations* that configure the objective categories therein portrayed. Unlike a street map, the brain's maps represent *abstract-feature domains* rather than concrete geographical domains. (Hence the allusion to Plato.) But as with maps generally, representation is achieved, not *atomistically* or one map-element at a time, but *holistically* or all map-elements together, by virtue of their collective internal structure, and by virtue of the homomorphism displayed between that internal structure on the one hand, and the similarity-structure of the relevant feature-domain on the other. The map is homomorphic with (at least a substructure of) the feature-domain being mapped. We might call this theory *Domain-Portrayal Semantics* to distinguish it from the various causal, covariational, indicator, teleological, and conceptual-role theories familiar to us from the contemporary philosophical tradition. Perhaps its closest cousin in that tradition is conceptual-role semantics (because both are holistic), but a contrasting feature of the present account is the fact that it has no dependence whatever on language-like structures and structure-sensitive inferences. It embraces all cognitive creatures on the planet, language-using or no.

These internal maps of sundry external feature-domains (e.g., voices, shapes, motions) are embodied in the high-dimensional activation-spaces of the brain's many distinct neuronal populations, populations that typically number in the tens of millions of neurons. And just as any specific point on a two-dimensional highway map is specified by the simultaneous values of two variables—latitude and longitude—so is a specific point in an *n*-dimensional neuronal map specified by the simultaneous values of *n* variables—the current activation or excitation values of each of the *n* neurons in the representing population. As the number *n* climbs beyond the two dimensions familiar from a street-map, the representational power of the *n*-D map climbs proportionately. With the number *n* presenting in excess of tens of millions, for each of perhaps a thousand distinct maps within the brain, all of them interacting with other, one starts to conceive a new respect for the representational powers of the biological brain, even for creatures well below us on the phylogenetic scale. As well, it now comes as no surprise that the bulk of one's background knowledge is deeply inarticulable.

If the story of the brain's grasp of the world's background structure is vaguely Platonic (plus or minus a pre-birth visit to an abstract heaven), so also is the story of its unfolding grasp of the perceived here-and-now. Our perceptions of the ephemeral world are always and inevitably interpreted within the framework of whatever background maps we have already pieced together. Our perceptions make sense only against the background of our antecedently grasped concepts. For the primary function of our several sensory systems is continually to index *where*, in the space of abstract possibilities already comprehended, one's current empirical position resides. Our assembled sensory inputs, at any given moment, serve to activate a specific *pattern* of activation-levels across each of our waiting neuronal maps, a unique pattern for each map (remember: each map has its own abstract subject matter), a pattern that constitutes a ''you-are-*here*-pointer'' to a specific possibility among the many background possibilities chronically portrayed in that map. We might call this the *Map-Indexing Theory of Perception.*

Very well, but a central problem for philosophy has always been, ''How do we *acquire* our general knowledge of the world's categorical and causal structure?'' Putting nativism aside—both Plato's and Descartes'—we are left with a variety of empiricist stories that appeal to induction, hypothetico-deduction, falsification, Bayesian updating, or some combination thereof. But these are all ''category-dependent'' forms of learning. They all require a determinate conceptual framework already in place, within which hypotheses can be framed, data can be expressed, and empirical reasoning can proceed. How such background frameworks are acquired in the first place is left unaddressed. Lockean/Humean stories concerning simple impressions and their residual copies—simple ideas—do attempt to fill this gap, but such stories are not empirically plausible, neither in their account of how ''complex'' ideas are subsequently generated therefrom, nor in their account of how the alleged ''simple'' ideas were generated in the first place.

If we ask, instead, how the *brain* develops its manifold maps of various abstract feature domains, developmental neuroscience already holds out the sketch of an answer. *Hebbian learning* is a mindless, sub-conceptual process that continually adjusts the strengths or ''weights'' of the trillions of synaptic connections that intervene between one neuronal population and another, the very connections whose assembled weights *determine* the complex landscape of prototype-regions that constitutes the abstract map embodied in the

receiving population. Modify the synaptic weights and you modify the map.

More importantly, the Hebbian process of weight-adjustment is systematically sensitive to *temporal coincidences* among the many axonal messages arriving, from an upstream population, to a given neuron in the receiving population. Specifically, if a cadre of connections, a subset among the great many connections to a given neuron, repeatedly bring their individual messages to the neuron *all at the same time*, then the weight of each connection in that united cadre is made progressively stronger. As neuroscience undergraduates are taught, ''Neurons that fire together, slowly wire together.'' The receiving neuron thus gradually becomes a reliable indicator of whatever external feature it was that prompted the simultaneous activation of the relevant neurons in the sending population, the neurons whose axon-tips embody the connections at issue. Moreover, since the salient features in any environment are those that display a repeated pattern of development over time (i.e., a distinct causal profile), the unfolding behavior of our Hebb-instructed receiving neuron over time can become an equally reliable indicator of a salient causal process out there in the world.

This sketch puts too much weight, perhaps, on the importance of a single neuron. Remember, there are thousands, even millions of other neurons in the same population, which are presumably becoming sensitive, each in their own way, to some aspect or dimension of the same external feature-unfolding-in-time. It is the Hebb-trained population *as a whole* that eventually gains the important grasp of that target, and of the ways in which it contrasts with, or is similar to, a variety of other prototypical features-unfolding-in-time. In this way, presumably, does the mindless process of Hebbian weight-adjustment gradually produce an internal map of an entire domain of abstract features, even if the infant creature's synaptic connections start off with random weight-values. The objective statistics of our sensory inputs over time sculpt an internal representation of those statistics. That is, they sculpt a *map* of the world's chronic or enduring structure, both categorical and causal.

Thus does any creature acquire the skills of perception and causal recognition: it learns to activate appropriate points and paths through its background neuronal activation spaces. Much the same process subserves its acquisition of bodily motor skills and the skills of manipulating its physical environment, as opposed to just passively observing it. Here, too, Hebbian learning sculpts representations: representations of the space of possible *actions*. *Practical* wisdom, it emerges, has the same sort of neuronal basis as does factual or theoretical wisdom, and in neither case do ''laws'' (in the latter case) or ''maxims'' (in the former case) play any fundamental role at all. Instead, one's level of wisdom is measured by the accuracy and the penetration of the high-dimensional maps one has constructed for the relevant abstract domains, both factual and practical. Plato, once again, would be pleased.

This holds for one's perceptual and navigational skills in the social and moral domains, no less than in the various physical domains. Conventional wisdom has long modeled our internal cognitive processes, quite wrongly, as just an inner version of the public arguments and justifications that we learn, as children, to construct and evaluate in the social space of the dinner table and the marketplace. Those social activities are of vital importance to our collective commerce, both social and intellectual, but they are an evolutionary novelty, unreflected in the brain's basic modes of decision-making. These have a different dynamics, and a different kinematics, entirely.

Upon reflection, this should come as no surprise. Baboon troops, wolf packs, and lion prides all show penetrating social perception and intricate social reasoning on the part of their members. And yet, lacking language entirely, all of their cognitive activity must be fundamentally non-discursive. Why should humans, at bottom, be any different? Decision theorists, be advised. And moral philosophers. And jurists. And those whose job it is to study, and to try to repair, various cognitive and social pathologies. As with factual reasoning, practical reasoning and decision-making is something we have but barely begun to understand.

To return to factual reasoning, the nature of cutting-edge scientific research looks interestingly different from the neuronal perspective as well. Making theoretical progress emerges as a matter of finding ever more penetrating and successful *interpretations* of the antecedently interpreted empirical data. It is not (usually) a matter of constructing fundamentally new maps for interpreting nature—that Hebbian process takes far too long. Rather, it is a process of trying to redeploy our existing conceptual resources in empirical domains *outside* the domain in which those concepts were originally acquired. Accordingly, Huygens reinterprets light as an instance of traveling waves. Newton reinterprets the orbiting Moon as a flung stone. Torricelli reinterprets the atmosphere as an ocean of air. Bernoulli reinterprets a gas as a swarm of ballistic particles. Each of these reinterpretations brought new insights and novel predictions in its wake. Theoretical science emerges as the critical exploration of revealing *models* and profitable *metaphors*, a process that

involves the new use of old conceptual resources. Neural networks, as it happens, are entirely capable of modulating their normal conceptual response to any given class of stimuli. For the axonal projections that lead us stepwise *up* the brain's cognitive ladder(s) to ever more abstract maps embodied in ever more elevated neuronal populations, also project *downwards*, in many cases, so as to allow cognitive activities at higher levels of processing to affect the ways in which familiar sensory inputs get processed at lower levels of interpretation. Brains, in short, can steer the way(s) in which they interpret the world, by making *multiple use* of the concepts that the very different and much slower process of Hebbian learning originally produced in them.

These downward-flowing or *recurrent* axonal projections are important for any number of reasons, beyond the function just described. They are vital for producing prototypical *paths* (as opposed to mere points) in activation space, paths that represent causal processes-unfolding-in-time. And they are equally critical for mastering the recursive structures displayed in natural languages, for mastering the skills of arithmetic, the skills of geometry, the skills of logic, and the skills of music, all of which embody recursive or iterable procedures over well-formed structures. A brain with a purely feedforward architecture might do many things, but it could never master these skills. A brain with a recurrent architecture can.

Enough examples. We have gone through, or at least gestured toward, (1) a theory of concepts, with (2) an accompanying semantic theory; (3) a theory of perception, folded into (1) and (2); (4) a sub-conceptual theory of how any creature's conceptual resources are formed in the first place; (5) a sub-linguistic theory of motor knowledge and practical wisdom; (6) a sub-linguistic account of social and moral knowledge; (7) a sub-linguistic portrayal of practical reasoning and decision making; (8) a sub-discursive account of theoretical science; and (9) a non-Chomskean account of our mastery of language and other recursive activities. Plainly, we are looking at a unified theoretical approach with an unusually broad reach.

There is much more to talk about, especially about the surrounding matrix of *human culture* and the manifold ways in which individual neural networks—that is, you and me—depend on and interact with that most blessed matrix. It is not a matrix of illusion (as in the silly movie by that name), but a matrix of acquired wisdom, an active framework that embodies many of the best achievements of the many earlier brains who also swam briefly in its nourishing informational embrace. This observation serves to illustrate that the neurocomputational perspective here paraded is not a narrow perspective, focused exclusively on the micro-arcana of individual brains. On the contrary, it is a multi-scaled perspective that may finally allow us to construct a unified, and unblinkered, account of human cognition as it unfolds over the centuries. At the very least, it offers a systematically *novel* approach to problems that have always been central to our discipline. Concerning its future success ... I live in hope, as always. But now the reader will have some understanding of why.

I close with an historical parallel whose presumptive lesson will be plain to everyone. Recall our attempts to understand the nature of life, and the many dimensions of health, prior to the achievements of modern biology: macroanatomy, cellular anatomy, metabolic and structural chemistry, physiology, immunology, protein synthesis, hematology, molecular genetics, oncology, and so forth and so on. Those pre-modern attempts, we can all agree, were pitiful, as were the medical practices based on them. But why should we expect our understanding of the nature of cognition, and the many dimensions of rationality, to be any *less* pitiful, prior to our making comparable achievements in penetrating the structure and the activities of the biological brain? Where should philosophy go from here? The answer could hardly be more obvious: into the brain.