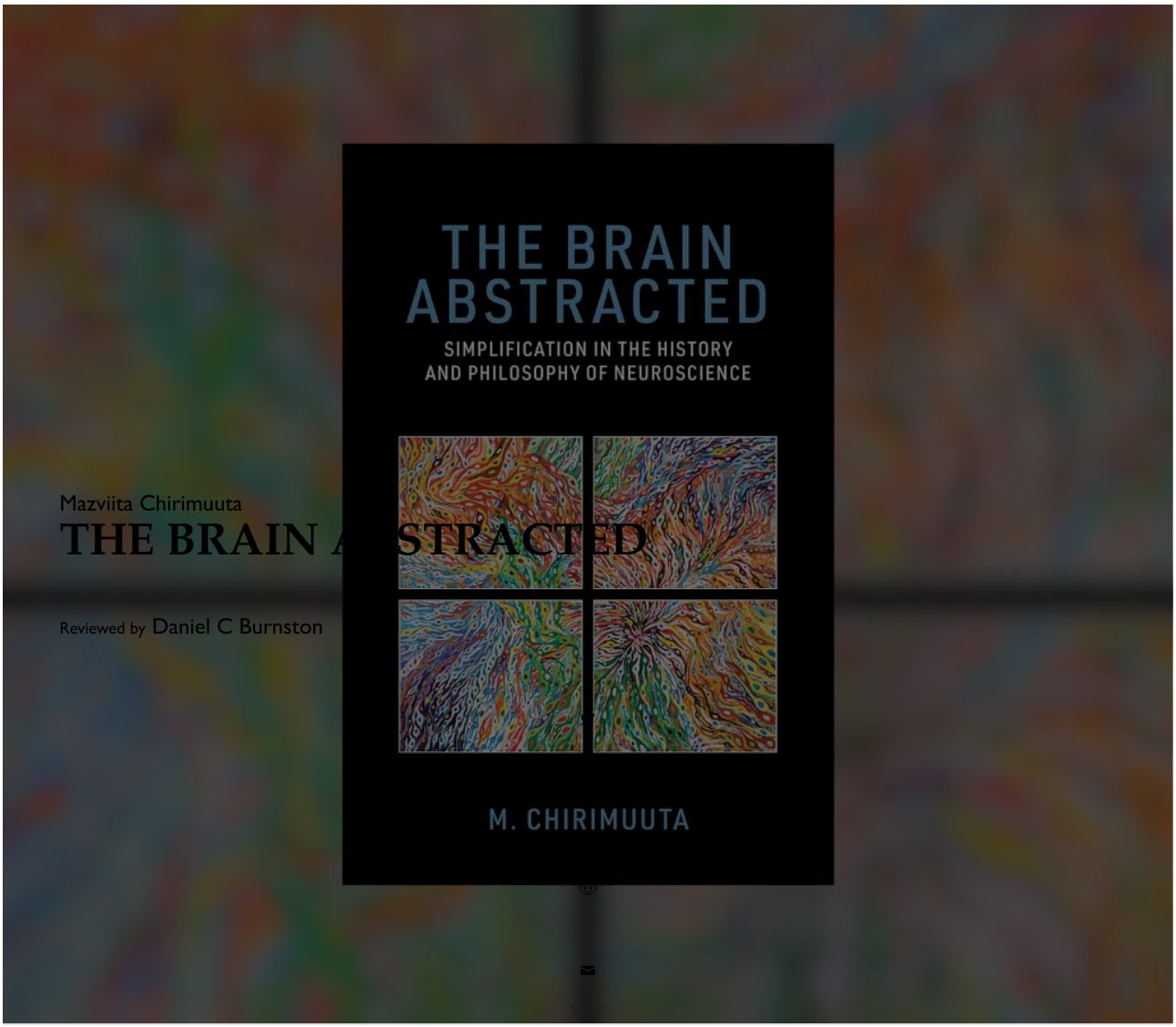

THE BRAIN ABSTRACTED

Mazviita Chirimuuta

Reviewed by
Daniel C Burnston

The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience
Mazviita Chirimuuta
Cambridge, MA: MIT Press, 2024, £48.00
ISBN 9780262548045

Cite as:
Burnston, D. C. [2024]: 'Mazviita Chirimuuta's The Brain Abstracted', *BJPS Review of Books*, **2024**, https://doi.org/10.59350/sprmp-kbw58

Mazviita Chirimuuta's *The Brain Abstracted* is a landmark work in the philosophy of neuroscience. The book offers an ecumenical view of neuroscientific practice, explaining both historical theories and cutting-edge approaches under a general philosophical framework. It gives a fresh take on long-running debates about

neural representation, computation, and functionalism, while also advancing the state of play within general philosophy of science. In short, it is an impressive book that is sure to shape discussion in the field for years to come.

The book is structured into three parts. In the first part (chapters 1–2), Chirimuuta gives a general philosophical framework with which to approach modelling perspectives in neuroscience. Part 2 (chapters 3–7) applies the framework to several detailed case studies from the history of neuroscience. Finally, part 3 (chapters 8–10) applies lessons from the first parts to ongoing debates in both philosophy and neuroscience. In this review, I will begin by outlining the contributions in each of the three parts, with specific focus on the strengths of the account. I will then give some criticisms of the meta-scientific approach in the book. The goal here is not to criticize the book writ large, but instead to highlight potential debates within the generally productive stance that it lays out.

Let's begin with part 1. According to Chirimuuta, neuroscience encompasses a diverse set of models, experimental approaches, measuring tools, and applications of new technology. What it doesn't involve is literal, direct descriptions of how the brain works. The brain is simply too complicated for that. The brain is structurally, functionally, and temporally heterogeneous. Among its billions of neurons, no two are exactly the same. Between those units, there are innumerable interactions and forms of organization. And the brain changes continually—plasticity and adaptability, rather than stasis and regularity, are the rule.

In theorizing about a system this complex, the best we will be able to get is a range of simplifications. In Chirimuuta's view, neuroscientific practice is organized around these simplifications—they are the key analogies or ideas that structure investigation. I will refer to a set of practices organized around key simplifications as a 'framework' (although I don't mean anything technical by this). Chirimuuta's aim is to specify what might be called the 'content' of these frameworks, that is, which abstractions and idealizations they employ, and thus how they represent the brain.

Chirimuuta's philosophical view is a combination of haptic realism, formal idealism, and perspectivism. Haptic realism suggests that while there are real systems in the world to which we apply our theoretical frameworks, those systems are in an important sense constructed by us. Neuroscientists don't study the full brain directly; instead, they study reduced preparations, removed from physiological context, learning history, and behaviour. They explore the brain the way we might explore an object by touch (hence the 'haptic'), understanding it by manipulating it for their investigative purposes. As a result, while neuroscientists do discover reproducible patterns, and apply theoretical frameworks to them, those patterns are not part of the natural furniture of the world. Formal idealism about these patterns denies the inference from successful modelling to ontological conclusions about the system we're studying.

Perspectivism follows directly. If the foregoing is correct, then there is no 'framework independent' way of generating scientific knowledge, and no 'real' brain that all neuroscientific frameworks converge onto. Since individual frameworks create their own, simplified version of a brain to study, they will be myopic, oblivious of genuine facts about the brain that fall outside the scope of their simplifying strategies. And since different frameworks involve different simplifications, they will involve 'unrecognizably different descriptions' (p. 46) of the brain. This has upshot for thinking about how they work relative to each other. For one thing, they are likely to remain dissociated from each other—this is a 'division of labor', rather than an 'integrative' view of pluralism (Burnston [2019]). Similarly, novel frameworks are likely to displace, rather than integrate with, previous ones.

There are many advantages to Chirimuuta's view. For one, philosophy of neuroscience, with some exceptions (Bechtel [2015]), has always had a decidedly realist bent. It has focused, primarily, on functionalist notions of psychological kinds and how they are realized in the brain. Chirimuuta's view is a timely and helpful corrective to this tendency. It just isn't the case that most psychological concepts or models map clearly onto discrete causal mechanisms in the head, and Chirimuuta's view explains nicely why this is so.

Relatedly, Chirimuuta's view points up the importance of thinking about disanalogies between one's model and the brain. Recognizing the inherent limitations of particular frameworks, she thinks, is one way that neuroscience advances. Meta-scientifically, we can use disanalogies to understand how science progresses—Chirimuuta suggests that, in Kuhnian fashion, new frameworks are often developed specifically to capture and explain the disanalogies between previous frameworks and the brain.

Lastly, Chirimuuta's view gives us resources to potentially diagnose, and avoid, ersatz debates. Many of these, Chirimuuta suggests, result from reifying distinct frameworks, and taking them as competing alternatives for the 'true' nature of the brain and cognition, rather than recognizing them for what they are: distinct simplifying assumptions, none of which are intrinsically more realistic than the others.

In part 2 of the book, the points above are illustrated through an impressive range of examples, which I can only briefly summarize here. They can be loosely classified as 'historical frameworks' and 'core concepts'.

There are three historical examples: reflex theory, simple cells, and the population vector view of motor cortex. Chirimuuta articulates the content of these frameworks via their shared commitment to reductionism. In a sense, reductionism is the meta-perspective that they share. Reductionism, on Chirimuuta's view, is an inherently atomistic perspective—one tries to describe the system's basic, unchanging components, and then explain the whole system in terms of their concerted operation. So, in the reflex theory, the idea was that behaviour (and the brain) could be decomposed into basic learned reflexes. The guiding thought behind simple cells was that the early visual system comprises a set of neurons with fixed spatial and featural receptive fields, which provide the input to more complicated feature extraction processes. Population vectors, in the motor cortex, similarly were taken to comprise cells that code for basic motor movements, which put together constitute a motor action.

Chirimuuta argues that these were the defining ideas of each framework, and further that critiques of these perspectives targeted their reductionist aspects. Opponents of reflex theory contended that it was a mistake to think of the whole brain in terms of the kind of simple stimulus–response reactions discoverable in the spinal cord, and that simple reactions like this could not capture the complex agency of the organism. Critics of the simple–complex cell hierarchy complained about these cells being discovered primarily in anesthetized or non-behaving animals. And critics of the population vector model argued that the basic unit is not individual cells but instead population-level activity. If Chirimuuta is right about these cases, then episodes in the history of neuroscience can be understood as the adoption of certain simplifying (for example, reductionist) assumptions, then critiques and eventual displacement of these assumptions by new frameworks.

The two core concepts that Chirimuuta analyses are the related notions of representation and computation. According to Chirimuuta, these notions are made up of distinct (if related) simplifying analogies that structure and enable investigation. In the case of representation, the analogy is to publicly available representational systems, which establish a semantic relationship between physically disconnected symbols and targets. This

analogy allows neuroscientists to focus on the causal relationships between brain activity and the distal world, rather than on the many proximate causal steps between them.

The guiding analogy in computational frameworks is that of information processing. The thought is that in the brain, like in a computer, only some of the many causal or physical processes are functionally relevant, particularly the ones that are performing transformations of information according to some algorithm. Thus, we can distinguish the causal interactions in the brain that are relevant for describing mental processes, from those background processes that provide resources and structural scaffolding for them. Describing the brain in this way also allows for formal mathematical principles to be brought to bear in describing brain activity.

Chirimuuta's goal here is to give 'charitable' but 'ontologically neutral' accounts of the core concepts—ones that capture their usefulness in science while also noting their limitations, and that don't make their use in neuroscience beholden to philosophical disputes (for example, about how to naturalize representation). The brain is disanalogous to a computer in many ways—it is not organized into algorithmic stages, for instance, and its physiology can only be partially mapped onto a mathematical function (her example is Marr and Ullman's famous computational view of 'edge cells' in the lateral geniculate nucleus). These disanalogies, however, do not necessarily undermine the perspective. In general, scientists need to 'strike a balance between pragmatic necessity of abstraction' (p. 115) and the false aspects of their models.

Finally, part 3 of the book is about ongoing debates. These include issues in both theoretical neuroscience and philosophy. The first concerns the advance of artificial neural nets as a model of the brain, particularly the visual system. The analogy of the visual system to a trained neural net has proven very powerful, as deep neural nets can be used to predict cell responses at multiple levels of the visual hierarchy. However, there is extensive debate about whether these models explain how the visual system works.

Chirimuuta gives a diagnosis of this dispute. One issue is that we don't have a view of the brain independently of a model or perspective, so we can't just 'look' at the brain to say whether artificial neural nets are good models of brain structure. Another is that, according to Chirimuuta, the models themselves are unintelligible. Their predictions are due to a huge number of small weight changes over extended training, and hence we cannot easily say what aspects of a deep net are producing its outputs. In a sense, we are stuck in the models we employ. Rather than trying to avoid this problem, Chirimuuta embraces it, even suggesting that we might revise the notion of 'understanding' to not require intelligibility of our model. It is sufficient, on this view, to have predictive knowledge of observable patterns.

In the final two chapters, which are among the more entertaining parts of the book, Chirimuuta takes aim at philosophical disputes that she thinks are the result of philosophers mistakenly reifying frameworks from neuroscience. One of these is the computer analogy. Theorists arguing in favour of machine consciousness based on the notion of 'functional isomorphism' between computers and brains simply mistake a kind of model (the computer) for its target (the brain). In doing so, these theorists ignore the disanalogies between brains and computers that may be (and, in Chirimuuta's view, likely are) relevant to understanding consciousness. Chirimuuta thinks that other long-running disputes, such as those surrounding scepticism and dualism, are the result of a mechanistic tendency to think of the mind as an internal part, fully dissociated from the body and the world.

Lastly, Chirimuuta argues that debates between representationalist and disjunctivist accounts of perceptual experience are ersatz ones, for similar reasons. Burge, for instance, endorses the framework of perceptual neuroscience, with its accounts of representation and internal processing, as a way of arguing against

disjunctivism. Theorists like McDowell, however, are after a different thing, namely, a philosophical account of how mind and world relate that rebuts Cartesian scepticism. Hence, a disjunctivist doesn't need to sign on for the framework that Burge reifies and can instead adopt their own framework for their own explanatory goals.

Chirimuuta makes clear that she is not giving a systematic, unified account of epistemology in neuroscience. Rather, she is giving an overall picture of the field, which she thinks has 'dialectical' advantages for understanding how it works. Hence, criticism of her view isn't appropriately pitched in terms of counterexamples or definitional disagreements. The question is whether her view does in fact describe how frameworks operate in neuroscience.

My criticisms in this vein can be summed up in two related points: the account, first, has an overly simplified view of the content of frameworks and, second, views those contents as more static than they are. In slogan form, frameworks in neuroscience are simplifications, but they aren't simple. Ironically, these problems point out that Chirimuuta's meta-science is itself a little reductive, trying to break frameworks down to their basic unchanging parts. The result is a kind of view where we have distinct ideas operating independently without much to say to each other, and no real guidance for how ideas can or should change over time, either in response to the data or to productive interaction with other theoretical ideas. Yet these things definitely do happen. As such, not illuminating how and why they occur, and why and when they are justified, is a problem for the account by its own argumentative lights.

These twin failures affect some of the particular accounts in the book, and I'll go through a few of these in brief. For one, the treatment of reductionism and mechanism in the book is unsatisfying. According to Chirimuuta, reductionism just is atomism, and mechanism is fully committed to the machine analogy—that is, that mechanisms must be composed of discrete, unchanging parts. This is despite the fact that a number of current mechanists, myself included (Burnston [2021]), simply deny that those aspects are constitutive of the approach. For these theorists, what is key is the ability to decompose the system. But decomposition can come in degrees and can be context sensitive, with mechanistic organization changing dynamically with context.

Chirimuuta's view seems to suggest that these views are just conceptually confused. I must say that I don't feel conceptually confused (maybe I'm wrong, of course…). And I have an alternative: Reductionism and mechanism are not just one idea. They have always comprised multiple related but dissociable components (the atomistic bits and the bits focused on decomposition, not to mention aspects of conceptual and theory reduction). Different mechanistic and reductionist projects can and do focus on different subsets of them over time. This would account for the fact that despite increased focus on context-sensitivity, system-level properties, networks, and so on, in current neuroscience, neuroscientists still often frame much of their activity in terms of the search for mechanisms.

A similar thing could be said about computation. Since Chirimuuta thinks computation is committed to a kind of algorithmic boxology, she doesn't give any attention to the idea of analogue computation. But neuroscientists these days are perfectly happy to talk about neural dynamics, state-space manifolds, and population trajectories as performing computations that explain behaviour. Concomitantly, philosophers have become interested in explicating notions of analogue computation that cover these kinds of explanations (Piccinini and Shagrir [2014]; Maley [2018]). Sometimes, ideas like 'computation' evolve and combine into new frameworks. Chirimuuta admits this in some cases: the idea of a reflex from classical conditioning theory has been taken up in modern reinforcement learning; the idea of feed-forward processing that inspired the notion of simple cells

also shows up in deep convolutional neural network frameworks, and so on. But these processes bespeak evolution and integration of frameworks as much as dissociation and displacement.

One would like a meta-science that captures both sides of this coin. I don't have space to pursue it at length, but there are some extant resources for this. There are, of course, integrative versions of pluralism, which take it as a norm that distinct frameworks should work in tandem wherever possible (Mitchell [2002]). There are also historical and philosophical analyses of concept change in the sciences, including case studies of 'genes' (Brigandt [2009]), 'cognition' (Colaço [2022]), and 'engrams' (Robins [2023]), as well as general 'patchwork' (Haueis [2024]) and 'open texture' (Makovec [forthcoming]) views of scientific concepts.

These concerns affect the assessment of the extant debates, too. With regards to deep neural nets, Chirimuuta seems to just assume that they are unintelligible, despite the many attempts to clarify how they might be profitably decomposed into more compressed descriptions, and how those could be mapped with varying degrees of directness to the brain (Cao and Yamins [2024]; Lillicrap and Kording [unpublished]). While mechanistic understanding of these systems is not guaranteed in the long run, it isn't something that can simply be ruled out either. Again, focusing too much on a simplified, fixed notion of the framework misleads about the state of the field.

With regards to philosophy, Chirimuuta's claim that the philosophical purposes of disjunctivism render it immune from science-based objections is, to me, discomfortingly open-ended. Does anyone who cites a philosophical aim—be it epistemological, ethical, decision-theoretic, or whatever—just inherently get out of ever having to worry about advances in the sciences? This would come as bad news to anyone who thinks that we better understand propositional attitudes, agency, rationality, perceptual representation, and other chestnuts by making philosophy productively interact with evolving scientific knowledge. The fact that sometimes this integration can go wrong doesn't mean it could never go right!

As a last criticism, I'd point out that there are times in the book where Chirimuuta does make seemingly straightforward ontological claims. She says, unqualifiedly, that the real brain could not be a computer, that there is no significant barrier between the brain and the world, that computers inherently could not be conscious, and so on. But it seems, by her own view, that these claims could only be made from an alternative framework. A more consistent application of the meta-science would state that simply identifying, say, computation and consciousness would be a mistake, but would resist the tendency to take another perspective that abstracts differently as 'more true'.

Let's sum up. I think that many of the philosophical positions in the book are extremely compelling, and that many of the meta-scientific points, diagnoses, and admonishments it makes are on the right track. I am less convinced that the overall meta-scientific angle in the book accurately captures the status of explanatory frameworks in neuroscience and philosophy.

I want to end, however, by stressing that my criticisms in fact point to the importance of the book. We can think of *The Brain Abstracted* as a kind of lodestar for philosophy of neuroscience. It captures and extends some of the major insights from recent philosophy of science, and applies them to the neurosciences. In doing so, it counterbalances the less reflective realist and functionalist tendencies in philosophy of neuroscience, and exemplifies a different way of thinking about how the field works. Like many broad schematics, it will be just as useful as a contrast for disagreements as it is for the many things it gets right. Everyone working in philosophy of neuroscience should read and benefit from it.

Daniel C Burnston
Tulane University
dburnsto@tulane.edu

References

Bechtel, W. [2015]: 'Can Mechanistic Explanation Be Reconciled with Scale-Free Constitution and Dynamics?', *Studies in History and Philosophy of Biological and Biomedical Sciences*, **53**, pp. 84–93.

Brigandt, I. [2009]: 'The Epistemic Goal of a Concept: Accounting for the Rationality of Semantic Change and Variation', *Synthese*, **177**, pp. 19–40.

Burnston, D. C. [2019]: 'Review of Angela Potochnik's Idealization and the Aims of Science', *Philosophy of Science*, **86**, pp. 577–83.

Burnston, D. C. [2021]: 'Getting over Atomism: Functional Decomposition in Complex Neural Systems', *British Journal for the Philosophy of Science*, **72**, pp. 743–72.

Cao, R. and Yamins, D. [2024]: 'Explanatory Models in Neuroscience, Part 1: Taking Mechanistic Abstraction Seriously', *Cognitive Systems Research*, available at <doi.org/10.1016/j.cogsys.2024.101244.>

Colaço, D. [2022]: 'Why Studying Plant Cognition Is Valuable, Even If Plants Aren't Cognitive', *Synthese*, **200**, available at <doi.org/10.1007/s11229-022-03869-7>.

Haueis, P. [2024]: 'A Generalized Patchwork to Scientific Concepts', *British Journal for the Philosophy of Science*, **75**, pp. 741–68.

Lillicrap, T. P. and Kording, K. P. [unpublished]: 'What Does It Mean to Understand a Neural Network?', available at <doi.org/10.48550/arXiv.1907.06374>.

Maley, C. J. [2018]: 'Toward Analog Neural Computation', *Minds and Machines*, **28**, pp. 77–91.

Makovec, D. [forthcoming]: 'Open Texture in Science and Philosophy', in E. Heinrich-Ramharter, A. Pichler, and F. Stadler (*eds*), *100 Years of Tractatus Logico-Philosophicus: 70 Years after Wittgenstein's Death*, Berlin: De Gruyter.

Mitchell, S. D. [2002]: 'Integrative Pluralism', *Biology and Philosophy*, **17**, pp. 55–70.

Piccinini, G. and Shagrir, O. [2014]: 'Foundations of Computational Neuroscience', *Current Opinion in Neurobiology*, **25**, pp. 25–30.

Robins, S. [2023]: 'The 21st Century Engram', *WIREs Cognitive Science*, **14**, available at <doi.org/10.1002/wcs.1653>.