DOI: 10.1017/psa.2025.26
This is a manuscript accepted for publication in *Philosophy of Science*.
This version may be subject to change during the production process.

Review of Cameron J. Buckner's *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*

Cameron J. Buckner, *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press.

Cameron Buckner's *From Deep Learning to Rational Machines* is an ambitious book that aims to show both that contemporary machine learning (ML) can meaningfully inform philosophy and that the history of philosophy can help guide the development of future ML systems. This is an exciting, timely, and valuable project. Furthermore, in service of these two goals, the book provides an accessible introduction to an extraordinarily successful approach to machine learning, termed *deep learning*. This will prepare readers to follow the main arguments in the book, as well as other philosophical work on machine learning, which is in itself a valuable contribution to the field.

The book is split into seven chapters, with the first two providing a very useful orientation on both contemporary ML and the nuances of the nativist–empiricist debate, and each of the last exploring a different cognitive faculty through the lens of a particular empiricist. Chapter 1 reviews the nativist–empiricist debate. Buckner's central contributions are (i) a recasting of the different positions in the debate with a more manageable continuum of theories, and (ii) a proposal for a modest empiricism of multiple domain general faculties. This proposal, which Buckner calls *Domain General Modular Architecture* (DoGMA), is inspired by medieval and early modern natural philosophers. While philosophers would benefit most from reading this chapter, it would also be valuable for ML researchers

who want to understand better the theoretical issues that lurk in the ambient intellectual atmosphere of their field. Chapter 2 surveys the modern state of deep learning and discusses some preliminary philosophical upshots. The chapter benefits from Buckner's rich experience as a philosopher and scientist of comparative psychology.

Each of the remaining chapters focuses on a particular empiricist faculty. While Buckner builds a coherent and cumulative narrative throughout the book, after reading the first two chapters the reader can choose to read only the particular chapters relevant to her interests. Chapter 3 connects modern image processing algorithms to John Locke's problem of mental abstractions; chapter 4 connects recent developments in reinforcement learning to Ibn Sina's account of memory; chapter 5 examines how generative deep learning models illuminate David Hume's conception of imagination; chapter 6 discusses how forms of William James' attention are implemented in specialized language processing algorithms; and chapter 7 explores how Adam Smith's and Sophie de Grouchy's developmental insights might improve deep learning models of social cognition.

Buckner's book shines in several ways. It provides the valuable service of explaining in an accessible manner many of the contemporary deep learning algorithms that underpin technologies like object detection, game solvers, and large language models. He then connects these algorithms to problems philosophers have traditionally cared about very deeply. These contributions dovetail with another success of the book: it sharpens certain concepts in the philosophy of cognitive science that significantly advance the field in a direction where arguments can be settled empirically. Chief among those concepts is the DoGMA. Some forms of empiricism, such as those exemplified in Richard Sutton's essay "The Bitter Lesson" (Sutton 2019), hold that only the most general learning algorithms—along with lots of data—are necessary for rationality. However, Buckner argues that the empiricist–nativist debate is a spectrum, with Sutton's views on one end and Jerry Fodor's on the other. The DoGMA is one view that lies on the more empiricist side of that continuum where instead of a simple, single master learning algorithm, multiple, different general learning algorithms working together can explain the cognitive feats of humans. Buckner argues that this is more in line with the empiricism adopted by its defenders in the medieval and early modern period, which he calls *origin empiricism* after the claim that there are no innate *ideas*, but instead innate *faculties*.

https://doi.org/10.1017/psa.2025.26 Published online by Cambridge University Press

The book also sheds light on debates in artificial intelligence (AI) about the degree of rationality we should impart to our machines by drawing upon lessons in comparative psychology. Some comparative psychologists avoid attributing complex mental abilities like theory of mind to animals based on behavioral tests, fearing overattribution errors. However, Buckner argues these researchers often exaggerate human performance on the same tests and overestimate the sophistication of human mental abilities. Humans frequently perform below ideal standards, and we lack certainty about whether humans use simple heuristics or sophisticated algorithms for abilities like theory of mind. Buckner suggests this same error leads skeptics to dismiss deep learning systems for supposed mistakes while overlooking actual human competence levels. His mastery of the relevant sciences, and associated philosophical issues, is on display in such discussions.

Although the book excels at showing how ML can inform the philosophy of mind, we find that the new DoGMA would have been more accessible to ML and cognitive science audiences, and more illuminating for philosophers, had it been more precisely formulated. Scientific evaluation requires assessing the truth of claims, and the DoGMA represents a substantive claim about cognitive architecture. Formalization within a common framework enables researchers to derive empirical predictions that can be tested, allowing for meaningful evaluation. While Buckner discusses some formalizations of competing nativist/empiricist positions, such as Gary Marcus' on p. 7, the DoGMA itself lacks such formal treatment. Either Buckner should have situated the DoGMA within existing frameworks like Marcus', allowing for direct comparison, or—more likely, given the nuance of Buckner's position—developed a novel formal framework that better accommodates it. Without formalization, it remains difficult to rigorously evaluate the DoGMA's applicability to AI systems. This absence is particularly unfortunate because we suspect Buckner's account offers greater subtlety and insight than competing formalized positions.

The core contributions we've described so far relate mostly to the first goal of the book: showing how contemporary ML can inform philosophy. In particular, the book does an excellent job at showing how machine learning can help adjudicate certain debates in philosophy of mind. This involves showing how the success of various ML systems can vindicate various empiricist ideas. However, the forward-looking, history of philosophy to ML direction is a bit murkier for us. Beyond broad brushstrokes, it is unclear to

https://doi.org/10.1017/psa.2025.26 Published online by Cambridge University Press

us in many of the chapters *which* particular ideas from the empiricists should be used to improve ML systems, especially since it is also clear that many of the details of the various accounts *don't* pan out. Furthermore, even focusing on broad lessons, we suspect that many of these have already been assimilated into the background perspective of ML, possibly undercutting how valuable the history of philosophy can be for ML.

For example, chapter 4 discusses Ibn Sina's account of memory. Buckner highlights that this account contains incorrect claims about brain structure and function, but writes that, if we ignore these, then we can see that "Ibn Sina offers a hierarchical, information-processing model where sensory information is processed and transformed into increasingly abstract formats" (pp. 159–60). It would be no news to ML researchers that looking for ways to deploy hierarchical information processing is a good strategy for designing sophisticated systems. Indeed, Buckner shows that recent developments in ML approaches to memory are inspired by such ideas in cognitive science, more detailed and plausible than Ibn Sina's. But this only highlights that perhaps we need to move beyond broad brushstrokes if, like cognitive science, the history of philosophy is to guide us in developing new ML systems.

Perhaps the book's most novel suggestions for ML researchers come in chapter 7, in which Buckner presents a new approach to enhancing ML moral reasoning through the ideas of Adam Smith and Sophie de Grouchy. In brief, the Smith and de Grouchy view is that moral sentiments are grounded in an empathic connection built through the dependence of young children on their caregivers. Buckner's suggestion is that AIs should be similarly constructed with moral sentiments for prosocial behavior. However, we still found that Buckner's proposal needed further fleshing out for it to be useful guidance. Thus, overall, the book would have benefited from an explicit discussion about how one might separate the chaff from the wheat regarding historical empiricist ideas.

The reader should keep a few things in mind when approaching this book. First, as discussed above, the book is focused on the philosophy of mind and cognitive science, and less on the epistemology of ML. For example, there is almost no discussion of statistical learning theory (the standard theory of learnability used in ML), nor explanations of why ML algorithms work in general, along with how we could use these insights to build better systems; instead, Buckner largely focuses on the *particular* architectures and their application to the nativist–empiricist debate. We believe that such topics deserve

https://doi.org/10.1017/psa.2025.26 Published online by Cambridge University Press

their own, fuller treatment than given in the book. Second, as mentioned above, the book is structured so that the reader can read specifically the chapters about the particular faculties that concern her. Given how carefully Buckner explores the connection between the philosophy, cognitive science, and the real-world ML architectures and problems, we think that each reader will get a lot of value out of the chapters that interest her, whatever her background. Third, due to the scope of the book, some of the ML topics are necessarily compressed in their presentation. We recommend readers treat the book as a solid introduction to the principles and approaches of modern ML, without being a replacement for a more detailed study of ML.

Buckner's book succeeds admirably in its first goal: demonstrating how ML can inform philosophical debates about mind and cognition. He makes a compelling case for ML's relevance to these fields, sharpens cognitive science concepts through insights from deep learning, and structures the book to serve both scientists and philosophers. However, regarding his second goal—showing how the history of philosophy can guide future ML development—the book falls short. Buckner provides insufficient guidance on which historical ideas merit adoption, which should be avoided, and how to distinguish between them. Moreover, many high-level philosophical insights have likely already filtered into ML through cognitive science and other pathways, making it unclear what additional value direct historical engagement might offer to ML practitioners beyond what has already been assimilated.

Bruce Rushing and Daniel A. Herrmann

References

Sutton, Richard. 2019. "The bitter lesson." www.incompleteideas.net/IncIdeas/BitterLesson.html.

https://doi.org/10.1017/psa.2025.26 Published online by Cambridge University Press