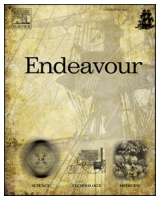
Endeavour 48 (2024) 100938

Contents lists available at ScienceDirect

Endeavour

journal homepage: www.elsevier.com/locate/endeavour



Book Review

A book review of From Deep Learning to Rational Machine: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence by Cameron J. Buckner. Oxford University Press, 2024, 440pp, ISBN: 9780197653302, £22.99. Hardback.

Cameron Buckner's *From Deep Learning to Rational Machines* defends a new dogma of empiricism. This "DoGMA" is the thesis that "an empiricist (Do)main General Modular Architecture is the best hope for modeling rational cognition in AI" (p. 30). The novelty of this thesis consists in combining insights from the history of empiricism in philosophy with technological advances from contemporary deep learning modelling. The goal is to show that computing systems built according to traditional empiricist principles can learn complex rational behaviours and acquire knowledge of external objects, other agents' mental operations and facts about norms and values, without the need for hardwired conceptual structures or mechanisms tailored to specific domains of cognition.

For Buckner, the successes of contemporary deep learning systems would also vindicate a moderate empiricist account of psychological development and rationality. Buckner surveys arguments and ideas about the workings of different faculties in human psychology by various authors in a broad empiricist tradition to frame foundational questions about the nature of the capacities of existing deep learning systems and to speculate about engineering innovations.

From Deep Learning to Rational Machines develops DoGMA over seven chapters. Chapters 1 and 2 set the stage for the rest of the book, by introducing the historical and contemporary debate between rationalists and empiricists and clarifying its relationship with distinct species of contemporary machine learning systems. Each of the remaining chapters zooms in on one psychological capacity, reviewing pertinent psychological insights from one author from the history of empiricism in philosophy and bringing them to bear on different processes and capacities displayed by deep learning systems. Specifically, Chapter 3 focuses on visual perceptual abstraction, deep convolutional neural networks, and John Locke. This chapter offers the strongest, most empirically adequate, case for DoGMA. Chapter 4 considers memory in relation to Ibn Sina and deep reinforcement learning systems that include modules for episodic control and offline experience replay. Chapter 5 engages with imagination, generative adversarial networks, and David Hume. Chapter 6 surveys William James' theory of attention and its incarnations in predictive systems and large language models with a transformer architecture that includes self-attention modules. Chapter 7 examines social cognition, theory of mind, empathy, and the emotions considering multidimensional deep reinforcement learning systems and Sophie de Grouchy.

One potential challenge for DoGMA, which Buckner acknowledges from the outset, is that contrapositions like empiricism vs. rationalism (and nurture vs. nature, learned vs. innate, …) are problematic and no actual author falls neatly into one camp or another. So, some readers might find the framing of contemporary debates about deep learning in terms of such crude dichotomies confusing.

But the type of *mindset* broadly shared within the contemporary deep learning community does tend towards an empiricist tradition, with its emphasis on training data, domain-general learning mechanisms, soft learning biases and domain-general non-conceptual, non-symbolic representations. Critics of deep learning sometimes argue that current deep learning systems cannot display genuine intelligence, since they lack innate knowledge databases and explicit operations over variables and symbolic representations with compositional structure – each an essential ingredient for human general intelligence according to a broadly rationalist tradition in linguistics and cognitive psychology.

At the very least, then, Buckner's framing helps readers appreciate that (casual) references made by some deep learning advocates and their critics to authors from the history of philosophy are often misleading. Actual empiricists have never argued for a completely "blank slate" picture of the human mind but emphasized how the extraction of abstract knowledge from sensory experiences recruits pre-existing, active, interacting psychological faculties, such as perception, memory, imagination, attention, and empathy. Buckner convincingly shows that this emphasis coheres with the architecture of successful deep learning systems, and with characterizing the dichotomy of empiricism vs rationalism in terms of domain-general vs. domain-specific learning mechanisms and representations.

Two more substantive problems for DoGMA are the empirical adequacy of deep learning models in accounting for *human* cognitive performance and the perspicuity of the notions of *domain-specificity/domain-generality*. The canonical authors surveyed by Buckner are interested in human or human-like intelligence.

But whether and how deep learning models provide us with empirically adequate understanding of the workings of the *human* mind is contentious. As Buckner acknowledges, the best case for deep learning as a model of human psychology relies on similarities between the architecture and processes in the mammalian visual cortex and in deep convolutional networks; but even here, there are remarkable differences between biological and artificial systems. The connection between other human cognitive faculties and deep learning architectures often remains suggestive rather than fleshed out (cf., Colombo & Ritchie 2024 for a recent Special Issue focused on these types of issues).

If it is contentious that successful deep learning models illuminate the workings of the human mind, it is also unclear how to define and identify domain-general mechanisms and representations. Generally, domain-specific mechanisms are said to be responsive to a narrow range of kinds of inputs (e.g., either visual or auditory inputs, but not both), while domain-general mechanisms would be responsive to a larger range of kinds of inputs (e.g., both visual and auditory inputs); domain-

https://doi.org/10.1016/j.endeavour.2024.100938

Available online 26 July 2024
0160-9327/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

specific representations would be tailored to be deployed in a narrow range of kinds of cognitive domains (e.g., either theory of mind or language, but not both), while domain-general representations would work across a wider range of cognitive domains. But these glosses are generic. And because the individuation of cognitive domains and inputs is far from obvious, it remains unclear how to count mechanisms and representations as domain-general or domain-specific, and in turn whether they are evidence for empiricism or rationalism. Additionally, similarly to other contemporary computational approaches to psychological change like Bayesian modelling (cf., e.g., Colombo 2018), deep learning modelling can incorporate both domain-general and domain-specific structures.

Despite these and other challenges, *From Deep Learning to Rational Machines* provides readers with a useful theoretical framework grounded in concepts and questions that philosophers and psychologists have traditionally been most interested in, a framework for evaluating the contribution of deep learning modelling in enhancing understanding of the character of (the human) mind, the nature of intelligence and rationality, and the workings of their underlying mechanisms. In the face of the hype, dread, and increasingly expansive roles of machine learning in our everyday lives, Buckner's *From Deep Learning to Rational Machines* is one of the first book-length treatments of deep learning by a philosopher, which will help researchers in multiple fields to get clear on foundational and methodological questions and disagreements in AI that also constitute a critical element in understanding human nature.

References

Colombo, M. (2018). Bayesian cognitive science, predictive brains, and the nativism debate. *Synthese, 195*, 4817–4838. https://doi.org/10.1007/s11229-017-1427-7

Colombo, M., & Ritchie, J. B. (2024). Foundations of Deep Learning. An introduction to the Special Issue. *Cognitive Systems Research, 2024*, Article 101246. https://doi.org/10.1016/j.cogsys.2024.101246

Gregor E. Bös, María Jimena Clavel Vázquez, Michael Cohen, Matteo Colombo
Tilburg University, Tilburg, The Netherlands