Blagoj Naumov 16182162

Buckner Book Review

A classic debate in the world of philosophy and science is the empiricism-nativism debate. How does knowledge and how do concepts originate? Does the mind begin as a blank slate and acquire knowledge through sensory experience as posited by empircisits such as John Locke? Or do we have certain innate concepts and ideas as posited by nativists such as Jerry Fodor? In philosophy of mind, this dispute frames how learning, abstraction, and intelligence are understood. When extended into aritifical intelligence, the same question arises. Do machines learn better through empiricist means or must they be equipped with some sort of innate cognitive principles? Cameron Buckner aims to build a bridge between all of the aforementioned ideas in his book 'From Deep Learning to Rational Machines'. Buckner argues that empiricism can guide the development of artificial intelligence and that recent empiricist artificial intelligence architecture can vindicate the largely forgotten argument in favour of empiricsm in human intelligence. The book defends what Buckner calls the ''new empiricist DoGMA" (Buckner, 2023), which is the claim that a Domain-General Modular Architecure (as opposed to a nativist Domain-Specific concept) best models rational cognition. Drawing on the likes of Locke, Hume, and James, Buckner contends that deep learning vindicates empiricism by showing that knowledge can emerge from experience through domain-general transformations. The book further positions itself between the strawman argument of blank-slate empircism and extre extreme nativism and thus reframes the empiricist-nativist debate as central to understand how machines and minds can learn, reason, and imagine.

The book's structure reflects Buckner's systematic approach to the subject. It start with two introductory chapters which frame and establish the empircist-nativist debate and clarify its relationship to contemporary machine learning. Buckner then devotes five further chapters to exploring specific human psychological faculties and compares and contrasts them to the same supposed faculties in machines, with the help of computer sceince, neuroscience, and psychology. Chapter Three exames visual perceptual abstraction through John Locke's philosophy and deep convolutional neural networks. Chapter Four investigates memory via Ibn Sina and Depp reinforcement learning systems. Chapter Five explores imagination through David Hume and generative adversarial networks. Chapter Six analyses attentian using William James' theory and transformer architectures. Finally, Chapter Seven addresses social cognition, empathy, and morality through Sophie de Grouchy's sentimentalist framework. Each chapter pairs a historical empiricist philosopher with a specific deep learning architecture, demonstrating how philosophical insights can guide AI development and how technological implementations can vindicate philosophical speculation.

Herein lies the first limitation of Buckner's book, not because of what it argues but because of what it omits. Buckner's projects focuses exclusively on whether deep learning can achieve rational congition but does not engage with the ethical, political, and social implications of the AI systems he analyses. This omission is relevant because as AI is getting more and more advanced and takes a hold of a larger portion of all of our lives, these questions come to the forefront of people's minds. The deep learning systems that Buckner celebrates are also technologies that are implicated in algorithmic bias, misinformation propagation, labour displacement, surveillance capitalism, and potential threats to the democratic process (Kreps and Kriner, 2023). On the other hand, the very traditions from which Buckner draws upon can be seen as being engaged with questions about knowledge's relationship to power and society. For instance, Sophie de Grouchy, featured in Chapter Seven, connected empiricist epistemology to egalitarian political philosophy (Simó, R.H.,2016) and when part of the book title is 'What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence", it is not a stretch to expect some sort of analysis on the ethical and political dimension relating to the future of artificial intelligence. Buckner's treatment, however, remains focused on cognitive architecture rather than how these historical thinkers might link to and inform contemporary debates about the future of AI. For exmaple, let us say deep learning systems acquire biased representations from biased training data (which directly relates to Buckner's idea on transformational abstraction). Then, if neural networks extract abstract knowledge through systematic transformations of sensory-like input, what happens if that input reflects certain (historical) injustices? How might moderate empiricsm inform efforts to create more equitable AI systems? These questions seem to emerge naturally from Buckner's framework, nevertheless Buckner's work on bridging the empircist-

nativist gap and adapting it to AI system is still relevant and important, even if from a purely (somewhat speculative) theoretical and scientific viewpoint.

As previously mentioned, at the heart of Buckner's project lies a careful reframing of the empiricism-natvism debate. So what does this reframing look like and how exactly does Buckner posit his argument? Rather than accepting the strawman dichotomy between blank slate empiricism and innate domain-specific modules, Buckner suggests what he calls 'moderate' or 'origin empircism' (Buckner, 2023). This position, he argues position better represents empiricists such as Locke and Hume, holds that while the mind's representational contents (its ideas or concepts) must be derived from experience and the psychological faculties and architectural features that extract knowledge from experience can be innately specified. The key question then becomes whether these cognitive systems are domain-general (empiricist) or domain-specific (nativist). A central idea to Buckner's moderate empiricism is what he calls the 'Transformation Principle" (Buckner, 2023). Rather than requiring that concepts merely reproduce sensory impressions, the Transformation Principle states that "the mind's simlpe concepts (or conceptions) are in their first appearance derived from systematic, domain-general transformations of sensory impression" (Buckner, 2023).

The chapter on imagination (Chapter Five), deserves special attention as it is from this point onwards in the book that Buckner relies on a 'proof of concept' argument (Buckner, 2023). Thus, if Buckner can successfully vindicate the empiricist dogma in the Imagination Chapter, he can more or less do it throughout the entirety of the book. Buckner positions this discussion as a debate between David Hume and Jerry Fodor, with contemporary generative deep learning models serving as empirical arbiters. Fodor challenges empircist accounts by arguing that imagination must perform essential cognitive functions such as synthesising tokens of categories for thought and interference and distinguishing causal from correlational relations (Buckner, 2023). While on the other hand, Hume never explains how imagination actually accomplishes these tasks and linkens this process to 'magic' (Buckner, 2023). Buckner's response involves demonstrating how various generative architectures realise the Humean argument. He examines General Adversarial Networks (GANs) and shows how they synthesise novel composite ideas through mechanism like vector interpolation in latent spaces. Moreover, variational encoders and transformer architectures demonstrate imagination's role in creating novel compositional representations. Adversarial Networks illustrate how systems can be modified to favour novelty, simulating human-like creativety. By showing that vector interpolation in high-dimensional latent spaces can generate novel composite representations in ways that blur the traditional distinction between interpolation and extrapolation, Buckner challenges Fodor's assumption that associationist mechanisms cannot achieve compositional productivity. It must be noted however, that even though these systems simulate human-like imagination, whether AI systems genuinely experience imagination in the same way humans do remains to be debated it. There is utility to be found in the idea that imagination requires embodiment and while Buckner defends and vindicates empiricism to an extent, this does not necessarily negate or dismiss the nativist view.

Throughout the book, Buckner makes a case that deep learning vindicates empiricism in important respects. The achievements of systems like AlexNet in visual recognition, AlphaGo in strategic gameplay, and large language models in lingustic tasks demonstate that domain-general learning mechanisms, given appropriate architectural builds and sufficient data, can simulate rational cognition without innate conceptual knowledge. The transformational abstraction demonstrated by deep convolutional neural networks shows how domain-general operations can solve the problem that troubled Locke: forming general category representations from particular, mutually inconsistent exemplars.

At the same time, Buckner's charity towards empircism should not take away what can be learned from nativist insights, particularly concerning AI development. While the book may succeed in vindicating empiricism as a viable path to artificial intelligence, it also inadvertedly dismisses nativism, or at least puts it to the side. The question remains whether empircism alone can provide the most effiicient or robust path. Nativist perspectives highlight several considerations that receive insufficient attention in Buckner's book.

Firstly, sample efficiency remains a challenge for deep learning systems. While Buckner acknowledges this limitation, noting that deep convelutional neural networks are 'date-hungry'

and exhibit ‘sample inefficiency’ (Buckner, 2023), the full implications need deeper analysis. Human children learn visual categories from remarkably few examples, suggesting that evolution may have endowed us with inductive biases or architectural constraints that dramatically accelerate learning in ways not fully captured by current domain-general architectures. The counter-argument for this would be that it is impossible to compare the processing power of the brain to that of a neural network and that the more advanced and more powerful machines become, the more they will match human performance levels. While this might be true, it does not negate the fact that it may just be more efficient to encode certain domain-specific or innate frameworks in AI architecture as this may be the solution to the sample inefficient problem.

Secondly, the brittlenss of deep learning systems or their tendency to fail catastrophically on adversarial examples or transfer tests that humans handle trivially suggests limits to pure domain-generality. While Buckner rebuts some criticism by noting that human cognition can also exhibit brittleness, the comparative brittleness of current AI systems still raises questions about whether or not additional architectural constraints inspired by nativist theories might enhance robustness. The debate between empircism and nativism need not be winner takes all kind of answer to the question at hand. Hybrid architectures incorporating both domain-general learning mechanisms with certain domain-specific biases might be the optimal path forward. It is also harsh criticising Buckner for this issue since he does not necessarily disagree or dismiss this premise, he simply seeks to vindicate empircism rather than put down nativism, however this remains an important distinction to be made when debating the topic at hand.

Thirdly, it can be argued that attention mechanisms and architectural innovations like transformers (discussed in Chapter 6) represent precisely the kind of domain-general innate faculties that moderate empiricism permits. However, the line between domain-general faculties and domain-specific faculties can be slightly blurry. The question of which architectural features should be considered part of domain-general learning machinery versus which represent domain-specific adaptations deserves more attention. The arugment seems to be in such a middle ground that it does not really dismiss nor vindicates either theory, as proponents on both sides can somewhat agree with what is being said.

Despite these limitations (some of which are admittedly harsh), ‘From Deep Leaning to Rational Machines’ does manage to bridge the empiricist-nativist human to AI gap. Buckner succeeds in demonstrating how historical empiricist philosophy can guide AI research and how contemporary deep learning acts as. Potential evidence for moderate empiricist speculation about cognitive architecture in humans. Even though the explanations provided can be supplemented with a nativist ready and serve to compliment the argument at hand, Buckner still manages to somewhat vindicate empircism and protect it from the dismissals brought on by the nativist dogma. At the same time, while domain-general learning mechanisms can probably accomplish more than what many nativists would have supposed, nativist readings about things like sample efficiency and robustness remain important for advancing AI. The most productive path forward likely involves creative synthesis rather than partisan allegiance to either camp. A positive takeway from this is that Buckner’s book invites further conversation. It demonstrates that deep learning and empiricist philosophy can mutually benefit one another which in turn opens space for additional interdisciplinary work. While we can find things that are missing, such as the ethical and political implications, these questions become urgent precisely because Buckner has shown that building genuinely capable AI systems along empiricist lines is no longer science fiction but engineering reality. ‘From Deep Learning to Rational Machines’ thus succeeds not only as philosophical analysis but as an invitation to deeper engagement with the profound questions artificial intelligence poses for human self-understanding and social organisation.

2009 words

Bibliography:

Buckner, Cameron J.(2023) *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence'.* New York, 2023; online edn Available at: https://doi.org/10.1093/oso/9780197653302.001.0001 [Accessed 24 Oct. 2025]

Kreps, S. and Kriner, D. (2023). How AI Threatens Democracy. [online] Journal of Democracy. Available at: https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/. [Accessed 24 Oct. 2025]

Simó, R.H. (2016). The Philosophy of Sophie de Grouchy: Theory of Knowledge, Ethical Theory, Political and Social Philosophy and Feminist Thought. [online] Academia.edu. Available at: https://www.academia.edu/23970401/The_Philosophy_of_Sophie_de_Grouchy_Theory_of_Knowledge_Ethical_Theory_Political_and_Social_Philosophy_and_Feminist_Thought [Accessed 24 Oct. 2025].