This is a draft of a chapter that has been accepted for publication by Oxford University Press in the forthcoming book:

***Deeply Rational Machines***

*What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*

by Cameron J. Buckner

due for publication in 2023

**Chapter 1**

**Moderate Empiricism and Machine Learning**

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism and lots of blank sheets. … Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed.
-Alan M. Turing (1950)

*1.1 Playing with fire? Nature vs. nurture for computer science*

In human inquiry, the introduction of a grand dichotomy—good vs. evil, mortal vs. divine, emotion vs. reason—can take on the vital importance, as well as the attendant danger, of the discovery of fire. While such dichotomies support qualitative shifts in the reach of our theorizing, they are often quickly taken for granted, perhaps too quickly, as an elemental force governing the world and our place within it. The distinction between nature and nurture stands as a prime example. This opposition has animated the human intellect for thousands of years, motivating the systematic exploration of competing styles of theory in nearly every academic discipline. We tend to have strong intuitions as to whether human knowledge is produced by turning inward to unpack our innate mental endowment or by turning outward to interpret the cipher of experience, and the energy provided by these intuitions has powered a variety of scientific and technological innovations. As with other Promethean bargains, however, such advances are bought at the expense of new and persistent dangers. Vigorously rubbing these opposing intuitions against one another can generate friction without illumination, causing theorists to pursue a research program long after its empirical prospects have grown cold, or to lose sight of the details of one another's views in a haze of misunderstanding and

exaggeration. And, of course, fires that grow too large can burn dangerously out of control. Lest we get singed, these distinctions must be continuously watched and carefully tended—particularly when a powerful new source of fuel is tossed into the flames.

We are now in the middle of just such a conflagration, and the new fuel source goes by the name of "deep learning." Indeed, funding and research for deep learning is currently blazing; as of 2023, every major tech company's marquee R&D group is focused on deep learning, with fierce bidding wars for top talent. Most issues of prestige publications like *Science* and *Nature* feature one of these groups' latest experiments. These publications report a series of transformative breakthroughs in artificial intelligence, including systems that can: recognize complex objects in natural photographs as well or better than humans; defeat human grandmasters in strategy games such as chess, Go, shoji, or Starcraft II; create novel pictures and bodies of text that are sometimes indistinguishable from those produced by humans; sift through the faintest radio echoes to discover new exoplanets orbiting stars thousands of light years away; crunch massive amounts of data generated by particle accelerators to try to find counterexamples to the Standard Model in physics; and predict how proteins will fold more accurately than human microbiologists who have devoted their lives to the task.[1]

In short, deep learning's current fortunes are white-hot; but, as with all systems of knowledge acquisition, our expectations of its continued prosperity are shaped by our views on the nature-nurture dichotomy. Deep learning's current status and future development are therefore meaningfully informed by philosophical positions, particularly those on offer in the historically grounded but ongoing debate between empiricists and nativists. At first blush, this debate concerns the origins of human knowledge: empiricists hold that knowledge is derived from sensory experience, whereas nativists tend to be rationalists who instead prize our capacity to reason—usually driven by an innate theory of the world's basic structure and/or of

[1] For details, see (Baldi, Sadowski, and Whiteson 2014; Brown et al. 2020; Chowdhery et al. 2022; Jumper et al. 2021; Krizhevsky, Sutskever, and Hinton 2012; Ramesh et al. 2022; Shallue and Vanderburg 2018; Silver et al. 2017; Vinyals et al. 2019).

rational minds—as the source of genuine knowledge.[2] When treated as an approach to artificial intelligence, deep learning is already identified as a nurture-favoring, empiricist style of theory, though I argue that its achievements vindicate a moderately empiricist approach to cognition that is more nuanced and resourceful than the empiricism typically surveyed in evaluations of deep learning's potential. This moderately empiricist approach, legitimated by an investigation of the historical origins of the philosophical debate in the work of influential empiricist philosophers and the application of their views to the relationship between machine learning models and the mind, suggests that today's achievements in deep learning substantially increase the plausibility that rational cognition can be achieved—and is achieved, in humans, many animals, and, if we hope to succeed, artificial agents—without the aid of the innate theories or concepts usually recommended by the opposed, nature-favoring, rationalist faction of theorists.

While empiricist and nativists theorists fight over the past, present, and future of deep learning systems development, the current enthusiasm for empiricism in engineering and business threatens to burn out of control—though this particular strain of empiricism sometimes draws oxygen from a simplistic understanding of the relationship between the successes of deep learning systems and the way that humans and animals actually solve problems. Research is moving so rapidly that an influential deep learning publication can receive 20,000 citations by the time it is only two or three years old—many of those while it is available only on a pre-print archive, meaning that it has not yet gone through the normal process of peer-review by other academics who could skeptically assess its claims. Meanwhile, leading nativists are going hoarse calling for the fire brigade. These nativists worry that deep learning is being applied to a wide range of problems without a firm understanding of how or why it works, and that the solutions discovered by deep learning agents are brittle and do not generalize to new situations as well as the strategies deployed by humans and animals. Depending upon whether you ask empiricists or nativists, deep learning systems can either already process input data so effectively that they are at least slightly conscious and on the verge of achieving escape velocity into world-spanning superintelligence, or they can do little more than bludgeon problems with

[2] To forestall confusion, the philosophical rationalism attributable to thinkers like Descartes, Leibniz, and Spinoza is not to be conflated with the new "rationalism" associated with blogs like LessWrong or Slate Star Codex, for which the traditional philosophical distinction is orthogonal.

massive amounts of statistics and linear algebra that can imitate the outward appearance of human intelligence but, because they lack the underlying structure provided by the human mind's innate startup software, will never capture even the most basic aspects of human mentality.

Although deep learning can be understood in purely technical terms outside the nature-nurture dichotomy, and hence outside the empiricist-nativist debate, it is difficult to assess its prospects as a route to artificial intelligence except through its light, with all its attendant prospects and perils. This debate of course has ancient historical origins, yet influential scientists frequently invoke its terms to explain and motivate their current views. For instance, in a front-page *Nature* article, a team from Google's DeepMind division pitched their AlphaZero system—which can easily defeat human grandmasters at the Chinese board game of Go, a game that is in some ways more complex than chess—as operating with a "tabula rasa" or blank slate algorithm (Silver et al. 2017). This empiricist metaphor entered the Western lexicon via Aristotle's *De Anima* (III, 429b-430a), which compares the human mind to the wax-covered tablets which the Greek academies used for notes; these tablets were "blanked" by heating them until the wax melted, smoothing the surface for re-use. The metaphor for the infant's mind became canonical through its repetition by a range of empiricist philosophers, from Aristotle's inheritors Ibn Sina (Avicenna) and St. Thomas Aquinas (the latter of which summarized it with the Peripatetic Maxim, which states that "*nihil est in intellectu quod non sit prius in sensu*" or "nothing in the mind which is not first in the senses"—*De Veritate* 2.3.19), to the Early Modern empiricists John Locke and David Hume, with whom the view is today most commonly associated.[3]

Deep learning enthusiasts are not the only ones to summon the history of philosophy in this context. Contemporary nativists have also been eager to align the current debate with historical positions. In his critique of the AlphaZero paper, for example, the nativist psychologist Gary Marcus associates Silver et al.'s blank slate language with the views of Locke, who wrote that "all ideas come from sensation or reflection" (*E*

[3] Other philosophical traditions also have views which appear recognizably empiricist by the standards of this debate; for example, some of the Yogācāra Buddhist philosophers like Dharmakīrti are identified as empiricist by interpreters (Powers 1994; Tillemans 2021) and some have even wondered whether Western empiricists like Hume were influenced by exposure to Buddhist philosophy (Gopnik 2009). Other commentators, however, view such trans-cultural linkages with skepticism (Conze 1963; Montalvo 1999). At any rate, a very interesting book similar to this one could be written by drawing upon the faculty psychology in these alternative traditions to interpret and guide the development of deep learning. I am grateful to Amit Chaturvadi for drawing my attention to these potential parallels.

II.ii.2). Marcus could just as well have linked it to Hume, who declared that "all our simple ideas in their first appearance are deriv'd from simple [sensory] impressions" (commonly referred to as his "Copy Principle" –*T* 1.1.1.7/4). Hume, however, is more frequently the target of Judea Pearl. One of the most influential living computer scientists and a frequent deep learning critic, Pearl has recently worried that deep learning theorists take as self-evident a "radical empiricism" according to which all knowledge "can be analyzed by examining patterns of conditional probabilities in the data" (2021).[4]

The history of philosophy certainly speaks to deep learning's achievements, but not in terms as simple as these interlocutors suggest. Where they see a stark dichotomy, Locke and Hume develop their keystone mantras into an elaborate empiricist theory of human cognition that is more nuanced and flexible. In fact, most research in deep learning is motivated by a set of assumptions more consistent with these philosophers' less radical take on empiricism, and one of the main tasks of this book is to articulate exactly which version of empiricism is most supported by recent developments. Identifying and clarifying the moderately empiricist approach too often lost in the flashpoint debates can unlock untapped explanatory power, both for understanding deep learning's current methods and for charting the optimal course to future breakthroughs. The challenge is that as with political slogans, even the seemingly simple statements of the empiricist creed can mean different things to different constituencies. By putting in the interpretive work to understand them charitably, we can avoid talking-past and direct evaluative efforts towards fruitful future research.

Unsurprisingly, even the most prominent nativists and empiricists today interpret the aforementioned slogans to imply quite different things. Nativist-leaning theorists tend to associate blank slates with the last great empiricist inferno, the behaviorist movement in American psychology, which reached the height of its power and then quickly dwindled to embers in the middle of the last century. Such theorists typically connect the empiricist blank slate with radically insufficient explanations for human learning. Steven Pinker articulates this perspective clearly in his book *The Blank Slate*. According to Pinker, today's empiricists have revived the

[4] In general, Pearl is less concerned here with the debate over nativism and anti-nativism in psychology than these other critics, and more engaged in the battle between skeptical Humean and realist approaches to causation in metaphysics and philosophy of science.

doomed mission of the behaviorists, who "through most of the 20th century…tried to explain all of human behavior by appealing to a couple of simple mechanisms of association and conditioning" (Pinker 2003).[5] Lake et al. also called out the "strong empiricism of modern connectionist models" which they identify with the "oversimplified behaviorism" that was "repudiated" by the cognitive revolution in the latter half of the 20th century (2017, p.4). This reported abrogation occurred when Noam Chomsky smothered behaviorism under a wave of his "Cartesian linguistics," which explicitly invoked the rationalist nativism of French philosopher René Descartes (Chomsky 1966) to inspire his arguments for an intricate set of innate grammatical rules to explain human linguistic ability.[6] Marcus even formalizes this behaviorist interpretation of empiricism by defining cognition as a function ranging over four variables:

$$\text{cognition} = f(a, r, k, e),$$

where $a$ = algorithms, $r$ = representational formats, $k$ = innate knowledge, and $e$ = experience. Marcus' construal of the empiricist approach—which, as mentioned above, Marcus attributes to Locke—"would set $k$ and $r$ to zero, set $a$ to some extremely minimal value, (e.g., an operation for adjusting weights relative to reinforcement signals), and leave the rest to experience" (Marcus 2018).[7]

On this point, nativists practice something of the radical simplification they critique, by assuming that for the mind to be "blank" at birth, it must begin with virtually no innate structure at all. The more charitable nativist philosophers Laurence and Margolis (2015) have recently worried that summarizing current debates in cognitive science as the question of whether the mind has any innate structure whatsoever has the unfortunate consequence that "there aren't really any empiricists."[8] In reality, a completely structureless mind, like an inert mineral slab, would not learn anything by being subjected to any amount of stimulus. This seems

[5] See also Childers, Hvorecký, and Meyer (2021), who also link deep learning to behaviorism; I defend a very different approach to linking deep learning to the history of the empiricist-rationalist debate.

[6] While also endorsing a rich package of "startup software" for the mind (which in their favored Bayesian models is typically programmed manually in symbolic form, including manually specified representational primitives and prior probability estimations) which they think should include components of Core Knowledge, Lake et al. (2017) are officially agnostic as to whether that software is innate or learned very early in childhood.

[7] What does "innate" mean here? An entire subarea of philosophy of science has burgeoned around the question of how best to define innateness (Ariew 1996; Griffiths and Machery 2008; Khalidi 2001, 2016, 2016; Mallon and Weinberg 2006; Mameli and Bateson 2006; Northcott and Piccinini 2018; Samuels 2004, 2007). For present purposes, we can proceed with a minimalist notion that implies at least "not learned" (Ritchie 2021).

[8] The empiricist-leaning developmental psychologist Linda Smith has also criticized this framing in her article, "Avoiding associations when it's behaviorism you really hate" (Smith 2000).

to be something that nearly all influential empiricists have acknowledged. Back in the twilight of behaviorism's reign, the empiricist philosopher Willard van Orman Quine observed that even the most radical behaviorists, like John Watson and B.F. Skinner, were "knowingly and cheerfully up to [their] neck in innate mechanisms" (quoted in Laurence and Margolis 2015; Quine 1969:95–96): they must assume a rich array of biological needs, sensory mechanisms, attentional biases, and reflexive behaviors which could be associated with one another before even the simplest forms of associative learning could begin. The items on this list suit organisms to their evolutionary niches without appeal to innate knowledge structures, illustrating why a more detailed examination of empiricist-branded theorizing in both philosophy and computer science is required. A more systematic examination of the history of empiricist theorizing quickly reveals appeals to innate factors more expansive than this list. Thus, while Marcus' formalized model of empiricism is sharper than the empiricist mantras in its implications, it is also less useful, particularly if we aim for a charitable evaluation of deep learning's prospects.

The preceding illustration of the empiricist-nativist dichotomy, as it informs the development of deep learning systems, offers a paradigmatic example of the nature-nurture dichotomy's enduring influence on human thought. Both distinctions are too often resolved into stark binaries, whereas the debate is better represented in terms of subtle continuums and differences amongst styles of explanation. Although the persistence of the opposition between nature and nurture suggests an unsolvable philosophical riddle at the heart of knowledge acquisition, it can, with care, be of use to us. The same is true of the empiricist-nativist dichotomy. When interpreted with more attention to the history of philosophy and its precise context of application, it can encourage more useful and principled debates between distinct research methodologies.

In fact, in cases where scientists have taken pains to understand the debate's history, it can be seen to have fostered notable scientific discoveries of the last century, such as Albert Einstein's theory of special relativity or the very invention of the digital computer and artificial neural networks over which today's debates rage. The philosopher of science John Norton argues that Einstein's theory of special relativity was inspired by his participation in a reading group on Hume's *Treatise* around 1902-1903 with the mathematician Conrad Habicht and philosopher Maurice Solovine, from which Einstein obtained a deep regard for Hume's

empiricism. In autobiographical notes from 1946, Einstein writes of his discovery of the relativity of simultaneity (to an inertial frame of reference) which undergirds special relativity that "this central point was decisively furthered, in my case, by the reading of David Hume's and Ernst Mach's philosophical writings" (quoted in Norton 2010). While rationalist philosophers like Immanuel Kant thought that absolute simultaneity was necessarily entailed by our a priori conception of spacetime, Einstein reasoned that if even these bedrock concepts were learned from experience, then there might be exceptions to them in extreme conditions, such as when objects travel at velocities approaching the speed of light.

Equally momentous achievements can be attributed to scientists listening to the nativist muse; the neuroscientist Grace Lindsay recounts how the early neural network pioneers McCulloch and Pitts (1943) idolized the rationalist philosopher Gottfried Leibniz, who theorized that the mind operates over an innate logical calculus from which all true propositions could be mechanically deduced (Lindsay 2021 Ch. 3). McCulloch and Pitts' idea that these complex logical and mathematical operations could be computed by large numbers of simple components organized in the right kind of network arrangement served as direct inspiration for both John von Neumann (1993) and Frank Rosenblatt (1958), whose works can be seen to have produced both the opposing research traditions responsible for the digital microprocessor architecture and deep neural networks (DNNs), respectively.

Here, I argue that the current incarnation of the nativist-empiricist debate in artificial intelligence presents us with a similar golden opportunity, in which we might attempt one of the rarest feats of intellectual alchemy: the conversion of a timeless philosophical riddle into a testable empirical question. For, if we could apply the distinction to the deep learning debate without confusion or caricature, then we could simply build some artificial agents according to nativist principles, and other artificial agents according to empiricist principles, and see which ones are ultimately the most successful or human-like. Specifically, we can manually program the nativist systems with innate abstract knowledge, and endow empiricist systems with general capacities to learn abstract knowledge from sensory experience, and compare the performance of the systems on a range of important tasks. Crucially, however, the empiricists in this competition must be allowed more raw materials than Marcus' formal specification allows, if we aim to hold a fair and informative competition.

If we could accomplish this conversion, philosophers and computer scientists would both reap the rewards. On the philosophy side, empiricists have frequently been accused of appealing to magic at critical points in their theories of rational cognition. Locke and Hume, for example often asserted that the mind performs some operation which allows it to extract some particular bit of abstract knowledge from experience but—given the scant understanding of the brain's operations available at the time—they could not explain how. Carefully examining the details of recent deep learning achievements might redeem some of the largest such promissory notes, by showing how physical systems built according to empiricist principles can actually perform these operations. Indexing the philosophical debate to these systems can further improve its clarity; where philosophical slogans are vague and subject to interpretation, computational models are precise, with all their assumptions exposed for philosophical scrutiny and empirical validation. Where successful, the plausibility of the empiricist approach to rational cognition substantially increases as a result. Of the benefits to computer science, philosophers have thought long and hard about the challenge of providing a complete approach to the human mind that is consistent with empiricist constraints, including how the mind's various components might interact and scale up to the highest forms of abstract knowledge and rational cognition. Deep learning is only now reaching for these heights in its modeling ambitions (e.g. Goyal and Bengio 2020), and so there may still yet be insights to mine from the history of empiricist philosophy that can be transmuted into the next engineering innovations.

To these ends, I here mount an interdisciplinary investigation into the prospects and implications of recent achievements in deep learning, combining insights from both computer science and philosophy. Doing so can both animate current engineering research with the warmth and wisdom of a classic philosophical debate, whilst simultaneously rendering the terms of that debate clearer than they have yet been in its long and distinguished history. Nevertheless, I know that such an interdisciplinary project is beset with its own distinctive risk. Richard Evans—an interdisciplinary researcher at DeepMind who has sought to create more powerful deep learning systems by augmenting them with logical maxims that he extracts from Kant's *Critique of Pure Reason* (including Kant's aforementioned maxim of simultaneity)—has issued a salutary warning for projects embarking under such ambitions:

> This is an interdisciplinary project and as such is in ever-present danger of falling between two stools, neither philosophically faithful to Kant's intentions nor contributing meaningfully to AI research. Kant himself provides: 'the warning not to carry on at the same time two jobs which are very distinct in the way they are to be handled, for each of which a special talent is perhaps required, and the combination of which in one person produces only bunglers.' [AK 4:388] The danger with an interdisciplinary project, part AI and part philosophy, is that both potential audiences are unsatisfied. (Evans 2020)

We must take Evans's (and Kant's) warning to heart. Yet, we must also acknowledge that, in part because deep learning is implicated in the nature-nurture distinction, philosophers are particularly suited to undertake the project. Whatever our other bungles, we have experience tending to this particular fire. To proceed, however, we must discard stools altogether. We will be better able to gauge the current and future achievements of deep learning by instead building a more accommodating bench, with ample room for a spectrum of distinctive backgrounds and expertise. Given the intensity of the current discussion amongst theorists grappling with deep learning's potential, the most productive way forward involves lowering the debate's temperature until the smoke clears, and inviting theorists from a variety of backgrounds with distinctive expertise and a stake in deep learning's implications to patiently work through the details together.

### *1.2 How to simmer things down: From Forms and slates to styles of learning*

Thanks to rigorous investigation in several disciplines, today we know that nearly all knowledge originates from a combination of both innate and experiential factors. Both radical nativism and radical empiricism are, in short, false. Despite this, more nuanced expositions of the distinction between empiricism and nativism remain the exception and have been almost entirely absent from discussions over deep learning.[9] Without a better way to understand the substance of the distinction, the recognition of this ecumenical outcome carries, on both sides, the threat of obliteration. This may be why Locke and Hume's empiricist mantras are so frequently interpreted as shoring up a diametric opposition between empiricism and

[9] The clearest examples one can find of such radical empiricism in the present debates are mostly found in grandstanding posts on social media sites; we should perhaps all let out a sigh of relief that figures like Descartes, Locke, and Berkeley did not have access to Twitter.

nativism. However, a closer examination of the history of empiricism suggests a more subtle continuum of views which can still support meaningful debates within the region of middle ground which remains empirically plausible. In fact, despite the allure of stark dichotomies, a historical review shows that empiricists, except for a few outliers, traditionally supposed that the mind begins with a significant amount of innate structure, and nativists typically acknowledged that most of the human mind's abstract knowledge is acquired through learning. We are more likely to generate insights relevant to both philosophy and computer science by wresting this moderation from extremism.

To begin, we can acknowledge that, contra the nativists reviewed above, the extremism of radical behaviorism was, from a historical perspective, just such an outlier. Rather than conflate all empiricist positions with behaviorism, we do well to follow further guidance provided by Laurence and Margolis (2012, 2015). According to their survey of historical and recent sources, it is more productive and widely applicable to construe the incarnations of the nativist-empiricist debate, both before and after behaviorism's heyday, in terms of two different styles of learning-based explanation:

> For contemporary theorists in philosophy and cognitive science, the disagreement revolves around the character of the innate psychological structures that underlie concept acquisition…According to empiricist approaches, there are few if any innate concepts and concept acquisition is, by and large, governed by a small number of innate general-purpose cognitive systems being repeatedly engaged….The nativist approach, in contrast, holds that innate concepts and/or innate special-purpose cognitive systems (of varying degrees of specialization) play a key role in conceptual development, alongside general-purpose cognitive systems. (2015)

Radical behaviorists proscribed theorizing about inner mental representations and faculties, because they worried that we could not provide objective empirical evidence for their existence. This proscription on theorizing about inner mental entities should be rejected, as it was both by the empiricist philosophy that came before it and by most deep learning theorists today. In fact, as glossed above, both sides today agree that the mind begins with a significant amount of powerful innate structure, that cognition involves complex interactions amongst internal representations and faculties, and that the vast majority of concepts are learned

or acquired from experience. The two sides still disagree, however, as to what exactly is innate. Nativists think that domain-specific abstractions can only be efficiently derived from a large number of innate, special-purpose concepts or learning systems that evolution has tailored to particular kinds of problems, whereas empiricists hold that general learning procedures can cooperate with a smaller set of domain-general cognitive resources to solve a wide array of problem types.[10]

Approaching matters this way places the weight of the distinction between the nativist and empiricist approaches on a corresponding distinction between domain-general and domain-specific cognitive systems or representations. The contrast might, in turn, be indexed to the range of inputs to which the innate resource responds: a highly domain-specific innate resource will be triggered only by a few very precise kinds of stimulus or situation, whereas a highly domain-general resource can be flexibly applied to a wide range of stimuli or domains. For instance, an innate prey-detection module that only controls tongue-darting movements and is triggered only by flies is more domain-specific than an innate memory store, which can record any arbitrary experience. Paradigm examples of domain-specific psychological constructs are the "innate releasing mechanisms" proposed by the ethologist Konrad Lorenz. He explained these systems using a lock-key metaphor: evolution prepares organisms via innate stimulus patterns (the "sign stimulus"), which are uniquely suited to unlock behavioral responses adapted to provide fitness benefits in response to just those stimuli (Lorenz 1935; Ronacher 2019). An all-purpose memory store, by contrast, can perform its storage and retrieval functions on any arbitrary input.

At issue in the distinction between domain-general and domain-specific cognitive systems is thus how inputs should be counted. Whereas the preceding tongue-darting example stands as a widely accepted illustrative starting point, we gain a better feel for the substance of the distinction by exploring a few specific nativist and empiricist views. Let us begin with nativism. Like empiricism, nativism can be understood as a continuum and manifests accordingly in more and less stringent forms. At the most radical end of the

[10] The debate, as it stands today, mostly concerns the learning systems of the human mind, or at least those of our closest primate relatives, and perhaps a few other successful evolutionary generalists whose cognition privileges power and flexibility, such as dolphins and corvids. Some less flexible animals with more predictable niches, like insects or amphibians, might have minds that are rigidly governed by just a few innate representations and behaviors, without this implying that empiricism about human knowledge or intelligence more generally is false or hopeless.

spectrum are the views of Plato and Jerry Fodor, according to which nearly every simple concept possessed by the mind is innate, including—as Fodor specifically adumbrated—the highly domain-specific concepts of CARBURETOR, BEATNIK, and QUARK (Fodor 1975; Laurence and Margolis 2015).[11] Further down the continuum, perhaps, come some of the most expansive drafts of Chomskyan linguistics, which posited potentially hundreds of innate grammatical principles and adjustable parameters thought to be common to all human languages—the "Universal Grammar"—to explain how human children efficiently home in on the precise grammar of their native language given limited experience (Chomsky 1986; Dąbrowska 2015; Lasnik and Lohndal 2010).[12] Most of these principles will only be activated by very specific kinds of grammatical structure, according to Chomsky, so this version of Universal Grammar appears to be a highly domain-specific, innate system.

Other nativists take inspiration from evolutionary psychology; Laurence and Margolis, for instance, sketch a still expansive but somewhat more biologically oriented list of special-purpose systems for dealing with "objects, physical causation, distance, movement, space, time, geometry, agency, goals, perception, emotions, thought, biological kinds, life stages, disease, tools, predators, prey, food, danger, sex, kinship, group membership, dominance, status, norms, morality, logic, and number" (Laurence & Margolis 2015). Laurence and Margolis think this list could implicate either innate concepts for each of these categories or innate learning modules specially tailored to acquire these concepts from experience. However, the key for

[11] I will here follow a common convention in this philosophical literature to put the names of concepts in smallcaps. Fodor's radical concept nativism is not actually as outrageous as it appears here, and he was probably trolling everyone to a significant degree in his bold claims that all these concepts are innate. All Fodor really means in saying that all these simple concepts are innate is that, if the mind has the kind of computational architecture he thinks it does, it needs to start with symbolic placeholders that could be triggered by the right experiences or linguistic labels, which later serve as pointers to files into which learning-based knowledge about the category can be collated. Later commentators have found ways to charitably salvage most of what Fodor says on the topic without endorsing his most absurd conclusions (Cowie 1998; Laurence and Margolis 2002; Sterelny 1989). As this is likely a prime example of a philosophical debate that has generated at least as much friction as illumination (and Fodor himself recanted this extreme concept nativism in later works, sort of—Fodor 2008), I will not comment further on it here.

[12] There is substantial variability in the estimate of exactly how many principles and parameters are required to specify UG in its most luxurious forms. In 1986, Chomsky suggested that it includes at least, "X-bar theory, binding theory, case theory, theta theory, bounding theory …[as well as] certain overriding principles such as the projection principle, FI (full interpretation), and the principles of licensing" (Chomsky 1986 p102). Fodor (2001) estimated about twenty parameters required to specify a grammar, and a wide variety of other estimates are available (Dąbrowska 2015).

them, and for other similarly situated nativists, is that the mind needs a rather elaborate and highly specialized array of start-up software to learn about these situations in a human-like way.

Other contemporary views that get labeled nativist tend toward the more domain-general side of the spectrum in their assumptions of innate endowments. Marcus himself puts in an order for a shorter list of more domain-general mechanisms than Laurence and Margolis, and the Core Knowledge developmental psychologists like Spelke and Carey—favorites of more moderate nativists in AI like Mitchell (2019) and François Chollet (Chollet 2019)—request only a handful of quite general innate concepts like OBJECT, AGENT, SOCIAL BEING, CAUSE, NUMBER, and SPACE (Carey and Spelke 1996; Spelke 1994). Regardless of how inputs are counted here, OBJECT and NUMBER almost surely apply to a wider range of stimuli and situations than DISEASE or KINSHIP, and should count as more domain-general. Later versions of the Chomskyan program—after his adoption of "Minimalism" around the early 1990s—aim to reduce the list of innate language-specific principles to perhaps only one, a rule for joining together different kinds of structured representations, called "Merge" (Berwick and Chomsky 2016; Chomsky 1993). Most recently, Chomsky seems to have made peace with the possibility that Merge might usefully apply to a range of non-linguistic domains as well, focusing instead on the conclusion that humans must have some kind of innate cognitive mechanism that allows us, and only us, to learn full-blown recursive language (Berwick and Chomsky 2017).

At this point, we can see that the gap between the nativist and empiricist positions has substantially narrowed—one referee's call of "slightly nativist" here could just as well be another's "mostly empiricist" there. But we should not let scorekeeping quibbles obscure the degree of consensus which has already emerged about the implausibility of the continuum's extremes. [13]

Turning to empiricism, it, too (as argued above) includes variations. However, empiricists generally aim to reduce the list of innate domain-specific learning mechanisms—and particularly those invoking innate concepts or ideas—into oblivion. Notably, the empiricist mantras of Locke and Hume referenced above only

[13] Pearl (2019) puts in perhaps the most modest requisition order of rationalist camp, insisting only upon rich mechanisms for representing and reasoning about the particularly broad domain of causation (some of which may be compatible with the domain-general approach to imagination that we will explore in later chapters).

concern the origins of the mind's initial ideas or concepts (and their representational contents). From context, it is obvious that when Locke talks about the mind being blank prior to experience, he refers only to its representational structures (its "ideas," in Lockean lingo), and not to its architecture or basic faculties. As the Stanford Encyclopedia of Philosophy entry on Locke ably puts it:

> While the mind may be a blank slate in regard to content, it is plain that Locke thinks we are born with a variety of faculties to receive and abilities to manipulate or process the content once we acquire it. Thus, for example, the mind can engage in three different types of action in putting simple ideas together … In addition to these abilities, there are such faculties as memory which allow for the storing of ideas (Uzgalis 2020).

Because it is necessary to distinguish this moderate empiricism from behaviorist construals, I hereafter refer to this doctrine as "origin empiricism." Importantly, origin empiricism does not prohibit the involvement of other active, innate factors in the process by which ideas are extracted from experience, or complex roles for previously-acquired ideas in guiding the abstraction process; indeed, both Locke and Hume liberally invoke innate, general-purpose faculties like memory, imagination, and reflection as well as complex interactions amongst previously-acquired ideas, even as they emphasize that the simple ideas from which these abstractions are crafted originate in sensations.[14]

In the context of AI, empiricist minimalism has been present since AI's prehistory, as evidenced from the quote by Turing with which this chapter began, which itself can be seen to channel the empiricist mantras of Locke and Hume. Yet, when nativists link historical empiricism to neural network modeling, they nonetheless tend to revert to reading empiricist figures as proto-behaviorists, despite behaviorism's aforementioned outlier status. This, in turn, leaves nativists stymied when empiricists appeal to innate faculties in their explanations of abstraction. For instance, when confronted by Hume's frequent appeal to an innate faculty of imagination, Fodor and Pylyshyn chide him for "cheating," because such faculties are something to which "qua associationist [he] had, of course, no right" (1988:p50, fn29). Such a charge saddles Hume with a tightly constrained view with little explanatory power; but Hume's Copy Principle, it is

[14] For a useful discussion, see Millican (2009).

important to note, says only that "simple ideas in their first appearance" must be derived from experience. It does not say that general-purpose faculties, like the imagination, must be learned, as well. So long as Hume's imagination can do its job without invoking innate ideas, Hume is only "cheating" if we insist that the rules of bridge require him to sit out this round, when the printed invitation he sent us clearly proposed a game of spades. To put this another way: It is perhaps Fodor's reading of Hume that is out of order, rather than Hume's appeals to the imagination.[15]

Whereas many nativists like Fodor tend to interpret mantras like the Copy Principle quite literally—as though concepts are just identical duplications of sensory impressions—more charitable readers like Laurence and Margolis permit empiricists to appeal to psychological processes that do more than merely reproduce impressions when manufacturing the mind's ideas or concepts. This is why I follow Laurence and Margolis in holding that the more productive way to construe the debate, particularly in the context of the dispute over the future of artificial intelligence, concerns only the number and domain-specificity of the innate concepts and learning systems that must be manually programmed into the mind to achieve human-like learning and cognition.

In fact, recent DNN models, which offer a standard toolbox of transformational operations that can do powerful computational work without invoking any explicit innate concepts, suggest an alternative way to construe the empiricist spirit present in Hume and others. To wit, the mind might more creatively massage and manipulate information originating in the senses in crafting the mind's concepts or ideas.[16] While I think

[15] For a systematic rebuttal of Fodor's reading of Hume, see Demeter (2021). For a book-length defense of the more constructive take on Hume's empiricism, see also Landy (2017).

[16] Recent empirical analyses of some of the most sophisticated deep learning architectures, transformers, suggest that there may be deep mathematical similarities between literal copying and some of the most abstract and impressive behaviors demonstrated by state of the art deep networks, such as translation. As a team from the startup at AnthropicAI (which includes several philosophers) put it in their analysis of transformers:

> "We emphasize again that the attention heads that we described above simultaneously implement both the abstract behaviors that we described […but] why do the same heads that inductively copy random text also exhibit these other behaviors? One hint is that these behaviors can be seen as "spiritually similar" to copying. Recall that where an induction head is defined as implementing a rule like [A][B] … [A] → [B], our empirically observed heads also do something like [A*][B*] … [A] → [B] where A* and B* are similar to A and B in some higher-level representation. There are several ways these similar behaviors could be connected. For example, note that the first behavior is a special case of the second, so perhaps induction heads are implementing a more general algorithm that reverts to the special case of copying when given a repeated sequence. Another possibility is that induction heads implement literal copying when they take a path through the residual stream that includes only them, but implement more abstract behaviors when they process the outputs of earlier layers

that empiricists like Locke and Hume all along supposed that empiricism was consistent with more than mere rote reproduction, for the sake of clarity and for the purpose of our investigation, I propose refining the Copy Principle, as the basic tenet of empiricism, into what I call the Transformation Principle:

> **Copy Principle:** "All our simple ideas in their first appearance are deriv'd from simple [sensory] impressions."
>
> **Transformation Principle:** The mind's simple concepts (or conceptions) are in their first appearance derived from systematic, domain-general transformations of sensory impressions.[17]

The Transformation Principle makes explicit that empiricist learning methods include a variety of default operations that can alter the structure of their input through systematic transformations, such as those studied in topology. Initially, the default methods should be domain-general, in the sense that they are not innately keyed to a specific set of features in a particular domain (such as faces, tribal affiliations, or biological categories). According to empiricism, such domain-specific transformations can be derived by specializing these generic operations later, through learning. This refinement prepares us to appreciate what I take to be the biggest lesson to be derived from deep learning's achievements: that a generic set of hierarchical transformation methods can be trained to overcome a wide range of problems that until recently most vexed progress in more nativist-inspired AI. Switching to an idiom of transformation also drives home that the basic toolkit of empiricist methods is computationally much more powerful than the forms of elemental stimulus-response learning favored by radical behaviorism, such as classical and operant conditioning.

Resetting the debate over AI in these terms can bring clarity to recent disputes over the empiricist implications of prominent models. Without such a reset, insights into deep learning developments understood through the lens of origin empiricism will continue to be frustrated. Such obstruction was the

---

that create more abstract representations (such as representations where the same word in English and French are embedded in the same place)" (Olsson et al. 2022).

[17] I include "conceptions" here to emphasize that the representations described by the Transformation Principle can be subjective in nature; two agents can have different conceptions of the same target category, and the same agent can change their conception of that category over time. In philosophy, the Fregean tradition of talking about concepts requires them to be objective entities properly studied by logic rather than psychology, and I am here intentionally avoiding that philosophical baggage (Buckner 2018; Gauker 2013; Machery 2009; Woodfield 1991). I do think there is good reason to maintain a tight relationship between subjective conceptions and objective reference, however; for a description of my full views on the relationship between conceptions and concepts in the context of representations in neural networks, see (Buckner 2022).

consequence of Marcus's allegation that Silver et al. overinterpreted the empiricist implications of their Go victory for the AlphaZero system (Silver et al. 2017). In support of his charge, Marcus noted specifically that 1) AlphaZero relied on Monte Carlo Tree Search (MCTS) to explore the implications of possible moves, and 2) some of its neural network parameters (such as the size of convolutional layers) were specifically tuned to the game before training began.[18] Yet, according to the preceding argument, exploring moves using MCTS is consistent with origin empiricism if the mechanism is equally applicable to a wide range of other domains (whereas configuring AlphaGo's convolution parameters specifically to the game of Go is indeed more suspect). In similar spirit, Marcus, in a public debate with deep learning pioneer Yann LeCun at NYU in 2017, called out LeCun's own "most famous" and "greatly-valuable" work on deep convolutional neural networks as inconsistent with his empiricism because it involves "innately-embedding translational invariance," or the assumption that important properties tend to be conserved across variation along systematic input dimensions like size, location, orientation, or duration (Marcus 2018). Marcus brandished LeCun et al.'s (1989) description of translational invariance as "a priori knowledge of the task" as a revealing gotcha moment. But bias toward translational invariance is clearly consistent with origin empiricism if it is built into the model so as to allow it to be applied by default to a variety of other stimuli and tasks in non-spatial domains as well.[19]

Marcus interprets deep learning's recent trend toward more complex architectures and inductive biases as a vindication of rationalist nativism, but even Marcus must admit that these general-purpose, structure-based biases are a far cry from the more numerous and specific kinds of innate representations that nativists have traditionally recommended. Descartes and Leibniz, for example, supposed that basically all of Euclidean geometry was contained within our innate ideas of space (Janiak 2020, Ch5). However, a DCNNs'

[18] In fairness, these parameters need to be set to some specific value, and it is not clear that the size of a convolutional layer can be translated into the kind of innate knowledge that rationalists have traditionally championed.
[19] Marcus emphasizes a point in this debate's Q&A exchange where LeCun seems to suggest that ideally, even translational invariance would be learned from experience. LeCun's point was likely that certain ways in which translational invariance is achieved in current DCNNs are biologically implausible and should be replaced by something with more fidelity to neural structure, if possible. In further clarificatory comments, LeCun endorsed local connectivity as a biologically plausible structural parameter which would likely not need to be learned in later incarnations of deep neural networks. Readers are encouraged to watch the entire Q&A after the debate if interested in this issue: https://www.youtube.com/watch?v=vdWPQ6iAkT4 .

bias towards translational invariance is at the same time more modest and more widely applicable than such innate geometry; it entails much less than the postulates of Euclidean geometry about the spatial domain, while at the same time applying much more broadly to any non-spatial data which exhibits the right kind of mathematical patterns.[20] It is crucial to acknowledge at this point that this is not a question of the original or primary evolutionary function of the bias; empiricists are free to agree that evolutionary considerations are important to understand the mind's organization—who today would deny that?—and that perhaps the demands of spatial reasoning posed an important fitness pressure that favored the brains of animals that imposed constraints on learning such as translational invariance. The central question for deep learning is rather how to best implement such biases in artificial systems.

The recent trend towards domain-general inductive biases thus reflects not the vindication of rationalist nativism, but rather the discovery that more flexible, domain-general, neurally inspired modeling methods can deliver dramatic performance gains on some of the same problems on which systems sporting domain-specific, manually encoded, symbolic versions of those assumptions had reliably and repeatedly failed. To understand this empirical and conceptual progress, it is more useful to focus on the way in which inductive biases are expressed in the mind. Nativists prefer to manually encode these inductive biases in a symbolic format that renders their interpretation transparent and constrains their application to a particular domain, whereas empiricists prefer to implement them using more global network structures that enable distinctive kinds of transformations—with extra points if those structures are "biologically plausible," with independent neuroanatomical support. The network theorists here emphasize that this latter approach applies the biases across a wider range of domains and situations, allowing domain-specific representations to emerge and connections to be drawn across very different kinds of inputs in potentially unpredictable and serendipitous ways. In short, domain-general inductive biases are critical to deep learning, regardless of philosophical allegiances. This explains why most deep learning theorists today not only do not hide their appeals to such domain-general inductive biases, as though they were embarrassing concessions to their

[20] Formally, it likely produces efficiency gains on any data patterns which exhibit a group-like structure (Achille and Soatto 2018; Bruna and Mallat 2011). Data patterns exhibiting group-like structure can be found far beyond the spatial domain.

nativist critics; they openly embrace them as the greatest achievements and future direction of their research program.

We see this in the dominant view at DeepMind—as expressed by senior figures like Demis Hassabis and Matthew Botvinick in the group's position papers (Botvinick et al. 2017; Hassabis et al. 2017)—which is more similar to the moderate origin empiricism recommended here.[21] In their response to Lake et al.'s (2017) criticisms, for example, they argue that AI should strive to design agents that can "learn and think for themselves" by eschewing "human hand engineering" and "representational structure" in favor of "larger-scale architectural and algorithmic factors" that allow agents to acquire domain-specific knowledge on their own (Botvinick et al. 2017). Yoshua Bengio—another central figure in deep learning—has endorsed the need for more complex neural network architectures featuring multiple structures that implement a set of domain-general inductive biases (Goyal and Bengio 2020 and see Table 1.1). Among the relevant questions today is thus not whether empiricists are allowed any innate structure or architecture whatsoever, but rather how much domain-specific knowledge can be extracted from incoming sensory experience using biased domain-general transformation methods, and how much instead requires domain-specific programming—whether by genes into brains, or by computer scientists into silicon—before human-like learning and experience can even begin.[22]

[21] James McClelland has also expressed this attitude in the work anticipating the current deep learning boom. Rogers and McClelland, for example, assert that "domain-general mechanisms can discover the sorts of domain-specific principles that are evident in the behavior of young children" (Rogers and McClelland 2004). He argue elsewhere that the transparent symbols and representations presumed by more nativist-friendly models are "abstractions that are sometimes useful but often misleading" and that the data they are adduced to explain can be better accounted for in terms of "generic constraints that foster the discovery of structure, whatever that structure might be, across a range of domains and content types" (McClelland et al. 2010).

[22] Marcus, for example, says that the neural substrates for innate, domain-specific representational content "are consistently localized across individuals, suggesting an important degree of genetic contribution to their neural organization"; and though they do not commit to the nativist option, Lake et al. (2017) specifically suggest that the mind may come "genetically programmed with mechanisms that amount to highly engineered cognitive representations or algorithms." Zador (2019) provides some arguments based on neuroscientific evidence as to why it is unlikely that any domain-specific knowledge could be genetically encoded in this way; it is much more likely that the genes specify more general wiring motifs that constrain the way neurons broadly self-organize, an approach which is consistent with the moderate empiricism defended in this book (for more discussion and examples, see Zaadnoordijk, Besold, and Cusack 2022).

| Inductive Bias | Corresponding Property |
|---|---|
| Distributed representations | Patterns of features |
| Convolution | Group Equivariance |
| Deep architectures | Complex functions are composition of simpler functions |
| Graph neural networks | Equivariance over entities and relations |
| Recurrent links | Equivariance over time |
| Soft attention | Equivariance over permutations |

Table 1.1 A list of domain-general inductive biases recommended by Goyal and Bengio (2020) that they think will allow DNN-based systems to extract domain-specific regularities across a wide variety of domains.

While it might appear that both sides have already met in the middle, we should not lose sight of the fact that they still draw inspiration from different sources and recommend different methodologies for future AI progress. For example, the critical structure behind DCNNs was inspired by neuroanatomical research rather than nativist developmental psychology (Hubel and Wiesel 1967), and it was first implemented by neural network modelers who at least saw themselves as motivated by the tenets of empiricism (Fukushima and Miyake 1982). Many contemporary nativists, moreover, still want much more specific and expansive innate knowledge, and emphasize entirely different sources of evidence. Evans, for example, aims to endow deep learning agents with the principles of Kant's a priori intuition of space, or the "subjective condition of sensibility," recommending that DNN-based systems be supplemented with a set of at least six formal axioms of Prolog that he derives from Kant's First Critique (Evans 2020, Ch. 6); and Marcus still officially recommends "hybrid" systems that combine neural networks with manually-programmed symbolic rules and knowledge structures extracted not from experience but from explicit theory in nativist developmental psychology. Thus, there remains a clear practical contrast in the research programs recommended by current empiricists and nativists, and deep learning's recent proclivity for general inductive biases is in no danger of sliding up a slippery slope to Platonic heaven.

### *1.3 From dichotomy to continuum*

Regardless, reconceiving the current nativist and empiricist positions according to the above allows for a more diverse spectrum of views to be arrayed along a continuum, and for individual theorists to decide which region of the continuum best captures the organization of the intelligent or rational mind, given current empirical evidence. To further aid the reconception, however, we must specify its most radical opposing points in the current debate. On the one side, the radical empiricism that worries Marcus and Pearl

finds expression in the "Bitter Lesson," a widely circulated blog post written by DeepMind-affiliated researcher Rich Sutton in 2019.[23] According to Sutton, 70 years of AI research should cause us all to worry that "the only thing that matters in the long run is the leveraging of computation" (Sutton 2019). According to this skeptical view, the programming of domain-specific knowledge that went into previous models was ultimately irrelevant and even counterproductive, as it hampered generalization to other tasks. The only thing that really improved performance in marquee AI achievements on this view, from Deep Blue to AlphaGo, was the application of more and more computation to larger and larger datasets. Even the largest deep learning models to date—GPT-3 with 175 billion parameters and a training set of over 40 billion tokens, or PaLM with 540 billion parameters—are thought by some to have fewer relevant parameters than the human brain, with its estimated 100 trillion adjustable synaptic connections (Hasson, Nastase, and Goldstein 2020). General intelligence will naturally arise, the student of the Bitter Lesson supposes, as soon as we stop trying to manually program human knowledge into machines in any form, and simply apply more data and computation to the solution of problems. This radical empiricism is fueled by recent results in "scaling research" on ever-larger networks and datasets; this research aims to show that qualitatively new forms of behavior can emerge by simple scaling up existing techniques to orders of magnitude more training or computation (Brown et al. 2020). Though I do not think this is the right interpretation of the Bitter Lesson (or even the lesson recommended by Sutton), this extreme empiricist interpretation—when combined with the extreme nativist interpretation on the other side of the spectrum, which might be indexed to the views of mid-career Chomsky or Fodor discussed above—will form a useful Scylla and Charybdis between which we can chart the development of the moderate origin empiricism defended in this book.[24]

When considering not only the continuum's radical terminal points, but also its continuity, we might note that the degree of representational structure that has been manually programmed into AI systems forms

[23] DeepMind is itself a big tent, employing figures from quasi-rationalists like Richard Evans, who recommends hybrid approach combining manually programmed Kantian assumptions about space, time, and causality with deep neural networks (Evans et al. 2020, 2021), to David Silver, who as we saw above invoked Locke in the attempt to eschew human knowledge entirely in mastering the game of Go.

[24] A better interpretation of the Bitter Lesson is that the right structural tweaks in machine learning should be those that reward increases in computation and representational resources with increases in performance. This has become regarded by many machine learning researchers as a kind of heuristic for deciding whether some architectural innovation is likely to pay off.

a central axis along which methods in AI have swung to and fro over the past seven decades. The 1960s and 1970s, under the direction of Chomsky and AI pioneers Allen Newell and Herbert Simon, the pendulum swung towards a high degree of manually programmed structure, finding its highest expression in the "expert systems" approach to artificial intelligence (Newell and Simon 1976). On this approach—which John Haugeland (1985) famously dubbed "GOFAI," for Good Old-Fashioned AI—progress was to be achieved by debriefing human experts in order to extract the abstract "heuristic" knowledge that allowed them to efficiently search for solutions to problems, and then manually program that knowledge into machines in the form of explicit rules and symbols (which, of course, they did not think was all innate; but they did not know how to create systems that could learn that heuristic knowledge for themselves).[25] IBM's Deep Blue, which bested Garry Kasparov in Chess in 1997 (Campbell 1999), and its Watson system, which defeated human champion Ken Jennings at Jeopardy in 2011 (Ferrucci et al. 2010) were products of this tradition[26], as is the Cyc system based at Cycorp in Austin, TX, which can be seen as its most dogged and enduring exemplar (Matuszek et al. 2006). Today, however, most see the methodological pendulum as having swung sharply back towards minimal domain-specific programming, after Watson was sold off in parts, and Cyc relegated to database translation and other jobs more modest than the grand ambitions under which it embarked. From this, we can align the continuum derived from the philosophical debate about the degree and domain-specificity of innate knowledge to the corresponding methodological debate in AI regarding the number and specificity of symbolic knowledge representation to manually program into artificial minds (Fig. 1.1).

[25] Some of Pearl's best-known research is in this tradition (Pearl 1984).
[26] Watson is known as a "hybrid" system in this context, as it integrates manually programmed symbolic knowledge with a variety of machine learning methods.

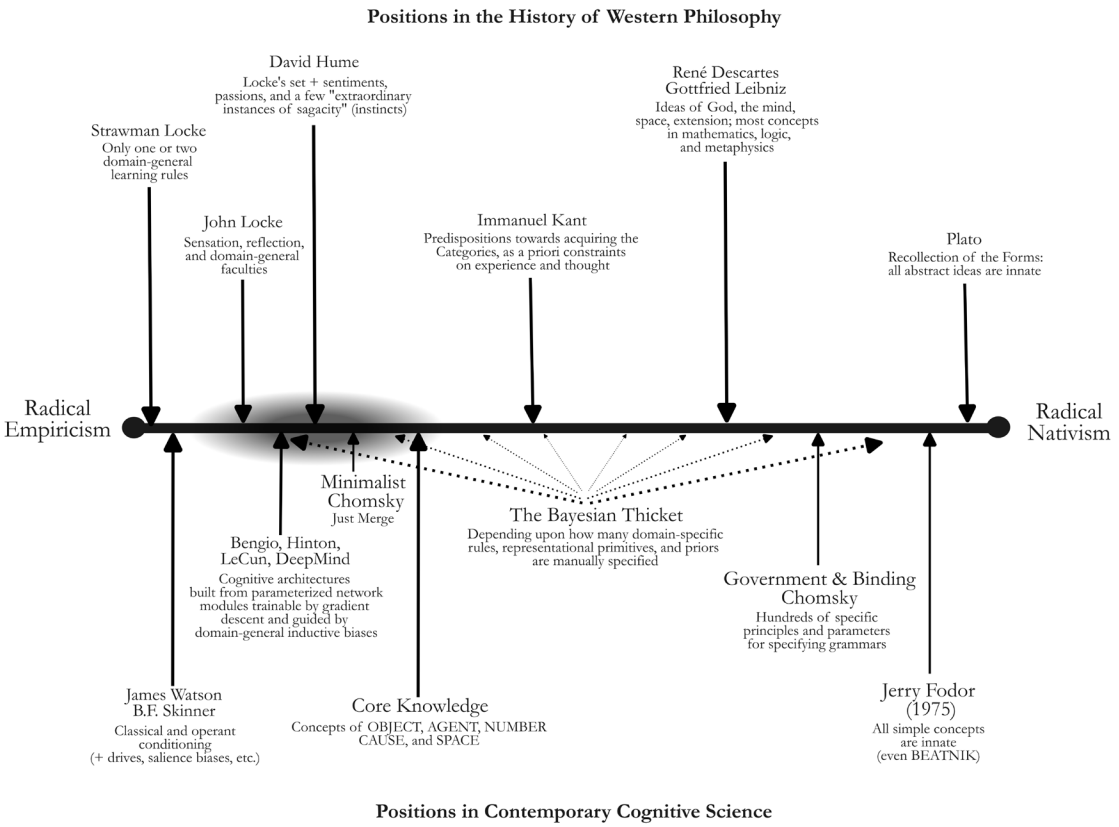


Figure 1.1 The distribution of various thinkers and positions from the history of philosophy and contemporary cognitive science according to the dimension of the number and domain-specificity of the innate ideas endorsed. The shaded left part of the continuum indicates the space of views which remain empirically plausible. It is hard to place Bayesian methods on such a continuum, as they can range from very domain-general to very domain-specific, with thousands of manually specified representational primitives and prior probability estimates.[27] This difficulty is highlighted with the "Bayesian Thicket," which indicates that they range from some of the most empiricist to some of the most hypernativist options on offer today. To decide whether a Bayesian model is more empiricist or more nativist, one must carefully scrutinize the selection of learning rules, representational primitives, and prior probability estimations that went into its construction. (The same is true of neural networks, of course, though the inefficiency of manually specifying parameters in neural networks makes this a generally unpopular choice.)

[27] The diagnosis of the generative linguist Norbert Hornstein of a Bayesian language-learning model in Perfors, Tenenbaum, and Regier (2011) is emblematic of this concern:

> "It is not clear how revelatory their conclusions are as their learning scenario assumes exactly the kind of richly structured domain specific innate hypothesis space the POS [poverty of the stimulus] generally aims to establish. So, if you are thinking that PTR gets you out from under rich domain specific innate structures, think again. Indeed if anything, PTR pack more into the innate hypothesis space than generativists typically do" (Hornstein 2012).

### *1.4 Of faculties and fairness: Introducing the new empiricist DoGMA*

To usefully investigate the similarities that unite and the differences that continue to divide moderate nativists and empiricists, it is necessary to rebut two specific aspects of the debate's common framing. First, I review the writings of the major empiricist philosophers, to emphasize their frequent appeals to faculty psychology. In other words, in eschewing innate ideas, these empiricists pushed the burden of deriving abstract concepts onto a set of active, general-purpose psychological faculties. Rational cognition on the empiricist model thus requires a cognitive architecture involving the cooperation of a variety of powerful innate faculties, such as perception, memory, imagination, attention, reflection, and sympathy/empathy. Not coincidentally, many recent breakthroughs in machine learning were obtained by adding computational modules implementing roles that philosophers and psychologists have attributed to one or another of these faculties. These interdisciplinary connections are often mentioned obliquely in computer science, but drawing them out and subjecting them to further scrutiny can provide further inspiration to machine learning modelers in the next round of innovation.

Reviewing these ideas from the history of philosophy may be particularly useful to engineers at the present inflection point. Recent faculty-inspired network models have mostly been "one-offs," focused on only a single faculty like memory or imagination. However, we know that a full cognitive architecture will involve many such faculties interacting with one another and competing for shared resources. The competition will create new problems for deep learning modelers as they aim to bootstrap their architectures to higher forms of rational cognition. While engineers often focus on the next benchmark or technical tweak, philosophers often consider the big picture before the details, and thus have elaborated rich ideas about empiricist cognitive architecture that anticipate problems engineers will face in the next steps of building an artificial mind.

Second, in the interest of facilitating more even-handed interdisciplinary communication, I occasionally provide a counterweight to enthusiastic summaries of evidence for domain-specific innate structure imported by nativists from other areas of cognitive science. For example, Marcus, Mitchell, Lake et al., and Chollet write as though the empirical case for innate Core Knowledge in very young infants is almost

incontrovertible. An even-handed appraisal would note that these positions are matters of active empirical debate in their respective disciplines, with a correspondingly distinguished list of empiricist developmental psychologists and biologists offering alternative—and I think often more compelling—accounts of the same empirical evidence. In many cases, I suggest, the nativists rely on an overly rosy or intellectualized view of the reliability, transparency, or robustness of distinctively human cognition and/or our current degree of empirical understanding of human uniqueness. In short, they often commit an error that I have dubbed "anthropofabulation", which combines anthropocentrism with a confabulated view of our own prowess (Buckner 2013). My empiricist counter-assessments often invoke evidence impugning the reliability of human judgments and deflating our current level of understanding of introspection and "common sense." In short, while deep learning models frequently underperform expectations, human cognition is also fraught with embarrassing errors, and we should not let our inaccurately high opinions of ourselves or double standards bias our evaluation of artificial agents.

To render it more memorable, I dub the moderate empiricism endorsed here the "new empiricist DoGMA." This acronym captures the attitude that an empiricist (Do)main General Modular Architecture is the best hope for modeling rational cognition in AI.[28] The acronym's lexical meaning is also intended to serve as a Sword of Damocles hanging over our heads, as a warning against hubris. The name should constantly remind empiricists that the DoGMA must be viewed as an empirical hypothesis which should be retained only as long as it remains consistent with the balance of empirical evidence and continues to support a fruitful research program.

The DoGMA also notably treats faculty psychology with a realist attitude that is sometimes considered off-limits to empiricism. This may be seen to go against a standard reading of the faculty psychology of, at least, Locke and Hume. Millican (2009), for example, suggests that the frequent appeals to faculties made by these empiricists are merely a "way of speaking," ultimately shorthand for more basic

[28] The acronym is a riff on the previous two dogmas of (logical) empiricism rejected by Quine—namely, the analytic-synthetic distinction (the idea that claims can be separated into those which are true by virtue of their meaning and those that are true by virtue of empirical facts) and semantic reductionism (the thesis that all empirically meaningful statements can be translated into sensory observation terms that would confirm or disconfirm them—Quine 1953).

associative operations. A famous passage in Locke that decries treating faculties as "so many distinct Agents" is often interpreted to support this view (*E* II.xxi.20). Demeter (2021) argues against this anti-realist reading of Hume's faculty psychology, suggesting that Hume intended to study the faculties from a third-person scientific perspective (in the way an anatomist would study internal organs), and that the comments from Locke and Hume which were interpreted as recommending an anti-realist attitude towards faculties were merely intended to contrast their approach to faculty psychology with the introspective approach to the faculties favored by the rationalists.[29] The empiricist reservations about faculty psychology, on Demeter's reading, stem from rationalists such as Descartes treating the faculties like introspectible sources of distinctive evidence and mental powers, rather than like theoretical posits whose operations can only be uncovered through empirical observation and reasoning. As Demeter (2021) puts it,

> Faculties thus conceived are not intuitively accessible causal sources or postulates of some preconceived hypothesis in the framework of which experience is to be interpreted; they are conclusions of experimental reasoning, and their identity depends on whether the analysis of relevant observations is correct (*T* 1.2.5.19). Instead of arguing *from* faculties, Hume argues *to* them; they are not the beginning but the aim of proper, experimental inquiry that reveals the characteristic activity of faculties.

[29] There are reasons to resist a simplistic approach to "mental anatomy" here, in that "mental organs" would likely need to share resources and interact with one another to a higher degree than physical organs. Neil van Leeuwen (2013:223) puts the worry in this way:

> "Perceptions, beliefs, emotions, mis-perceptions, rational inference systems, irrational biases, items in memory, etc. are all partial sources of novel ideas, so they are all potentially components of constructive imagination. Otherwise put, human imagination is built out of components, many of which also serve other purposes; it's not a single 'faculty.' The great ancient anatomist Galen discovered multiple biological purposes for every bone he analyzed; I think something similar holds for the building blocks of imagination: abilities such as memory and perception serve both reality tracking and constructive imagination (more on this below). Imagination, in part, is the capacity to use them for more than one purpose."

Nevertheless, creating artificial systems that can model such a capacity requires endowing them with subsystems which have distinctively generative computational operations; an architecture with a perceptual recognition and a memory storage system will never be able to model these generative capabilities without further supplementation. Indeed, it has long been a refrain of empiricist faculty psychology, all the way back to Aristotle and Ibn Sina, that the faculties exist in a kind of hierarchy and that "higher" faculties will draw upon and in turn down-regulate lower faculties; this does not imply that each faculty lacks distinctive computational operations, which is the focus of the moderate empiricist modeling methodology recommended in this book.

On this reading of Hume, the DoGMA is fully consistent and continuous with Hume's project. Faculties on this construal are like "mental organs" with distinctive operations to be uncovered by empirical reasoning, rather than by introspection or functional definition. The discovery of the critical aspects of those faculties that enable their distinctive operations—through DNN-based modeling—is a core aim of the DoGMA.

The DoGMA can be further characterized in terms of a negative and a positive side. The negative side prohibits appeal to innate ideas or concepts. The positive side commits us to derive domain-specific abstract knowledge from experience via the operation (and cooperation) of active cognitive faculties embodying domain-general inductive biases. As a "to-do" list for machine learning, the DoGMA recommends that we model these faculties using architectures featuring multiple, semi-independent neural network modules characterized by distinctive wiring motifs or learning algorithms. My use of the term "module" here should be explicitly distinguished from Fodor's use of the term (Fodor 1983), as Fodorian modules are characteristically domain-specific and strictly informationally encapsulated; though the modules of the DoGMA share other properties in common with Fodor's list, such as performing their own distinctive operations, exhibiting a greater degree of internal than external information-sharing, and possessing a fixed neural architecture.[30] These modules may, but need not, correspond to neuroanatomically distinct brain regions; indeed, we may expect that while these faculties perform distinct psychological functions, they compete for shared resources in the brain and their neural correlates might spatially overlap and even share many components.[31] Once articulated, we will find that the project of modeling distinct faculties with distinct

[30] I thus invoke here a particularly weak notion of a "module," even by permissive recent standards (Robbins 2009). The notion invoked here is closest to that recommended by Carruthers (2006), though he also emphasizes the domain-specificity that I here explicitly reject. I take no stance here on whether the mind is massively modular or also contains learned, domain-specific modules—I only here suggest that a useful way forward in deep learning research involves explicitly theorizing about computational models with distinctive operations that are more internally than externally connected and which implement roles attributed to domain-general psychological faculties in a coherent cognitive architecture.

[31] While I will introduce the faculties as distinct modules, we may decide upon reflection that one of the faculties can be reduced to a mode of or interaction amongst the others, or that the operations of two faculties are so intermingled that drawing a distinction between them would be artificial. Indeed, some theorists have argued that memory and imagination are merely two different aspects of the same faculty, or that attention is merely a mode of the other faculties (De Brigard 2014; Mole 2011). Debates about the exact taxonomy of the faculties could reflect a healthy maturation of empiricist theorizing without threatening the basic point that empiricists can appeal to domain-general mental operations which exceed the most basic forms of associative learning. The DoGMA does stand in opposition to more radical forms of empiricism which seek to eliminate traditional faculty taxonomies entirely, such as the eliminativism recommended by neuroscientist Luiz Pessoa (Pessoa, Medina, and Desfilis 2022). While Pessoa suggests that the standard list of faculties

neural network architectures is already well underway in deep learning research, and that it is this moderate empiricist DoGMA—rather than behaviorism—which has inspired and is correspondingly bolstered by its recent achievements.

Even nativists and more radical empiricists who ultimately reject the DoGMA may benefit by considering its charitable characterization and defense. In fact, nativist-leaning readers may find more points of agreement with prominent deep learning theorists than they would have otherwise expected. For one, they may find many ideas traditionally championed by nativists—such as modularity (albeit, alas, not the domain-specific or massive flavors championed by Fodor 1983), model-based learning (albeit, alas, without innate concepts), and compositionality (albeit, alas, without unlimited productivity)—vindicated by the DoGMA. For another, most nativists may happily borrow from more powerful learning algorithms and architectures to integrate the recent achievements of deep learning into their own theories (as recommended even by critics such as Marcus). Charitable consideration of the DoGMA may even help nativists marshal empirical evidence as to exactly which ideas or concepts must be supposed innate. Thus, hopefully even nativists will see the value of systematically engaging with the considered, charitably construed views of the most influential empiricists.

I aim to show how prominent empiricist approaches are more reasonable and powerful than many have supposed, and I will consider the book a success if readers leave feeling that it is an open empirical question whether innate domain-specific structures are required for human-like AI, rather than being obviously required. Questioning the strength of the evidence for prominent nativist hypotheses like Core Knowledge is one route to this goal, but in the interests of space and novelty, I here focus on illustrating how much cognitive flexibility can be derived from domain-general structure in recent deep learning models, especially since much of this evidence is new to the debate and the pace of research has been so blisteringly fast.

---

derives from folk psychology and lack convincing empirical support, I will here approach them more as solutions to computational problems faced by network-based approaches to rational cognition. It is an empirical discovery in computer science that network-based systems need mechanisms like distinct memory systems and attentional control to make decisions efficiently and flexibly in challenging environments.

Readers spoiling for a fight over the empirical support for nativism might be referred to a number of other works that critically evaluate it more directly (Christiansen and Chater 2016; Elman, Bates, and Johnson 1998; Griffiths and Tabery 2013; Mameli and Bateson 2006; McClelland et al. 2010). However, it must be acknowledged by all parties that psychology's replication crisis has hit its developmental division hard, and even some prominent nativists in psychology have (admirably) admitted that their best-known empirical results and standard experimental practices have recently been called into question by systematic replication failures (Kampis et al. 2020; Kulke et al. 2018; Salvadori et al. 2015). A reform movement is currently underway in developmental psychology in particular (e.g. Frank et al. 2017; Oakes 2017; Rubio-Fernández 2019; The ManyBabies Consortium 2020); nativists may want to wait for the dust to settle there before insisting on specific components of Core Knowledge in their critiques of deep learning.

### *1.5 Of models and minds*

To forestall some obvious worries about the aims of artificial intelligence research as understood here, it will help to clarify the nature of the purported relationship between deep learning models and the human mind. I will not here be assuming or arguing for the position that current deep learning models themselves possess important mental properties like intelligence, cognition, rationality, or consciousness. Our ability to draw conclusions about the innate endowment of the intelligent or rational mind by examining current neural network models is immediately impeded by the fact that these systems are very partial models of intelligence that are unlike us and our brains in most of their details (something which could be obscured by the current degree of hype). Some of these disanalogies are obvious; these systems lack bodies, developmental histories, personalities, and social networks. It would be confused to offer AlphaGo a medal for having beaten Lee Sedol, for it does not care whether it wins, and it has no friends or family with whom to celebrate even if it did. These systems are, at best, slivers of rational agents. Further disanalogies are revealed by an under-the-hood inspection. Human and animal brains—the biological engines driving the only deeply rational machines of which we are currently aware—are rich symphonies of chemical interactions, exhibiting neural dynamics at multiple timescales and built using other non-neural but potentially relevant components like glia and myelin. We do not know which of these features are essential for the brain's role in supporting our rich mental lives,

and the current generation of DNNs make little attempt to capture all this complexity. While DNNs can produce impressively adaptive behaviors, these are mere brush strokes of the behavioral flexibility that humans placed in similar situations could exhibit and the computational models underlying them paint only very impressionistic pictures of the psychological and neural mechanisms that enable such flexibility.

It would be too hasty, however, to conclude that such disanalogies render DNNs irrelevant to debates over the innate start-up software required for intelligence or rationality. At minimum, these models serve as proofs of concept that systems adhering to empiricist constraints can extract certain types of knowledge from certain types of input data. More substantively, if there are abstract correspondences between the structure of a neural network and the structure of the human brain or mind, then study of the computational model might reveal how humans actually do it, and more specifically that we actually do it without innate domain-specific knowledge.

The question of abstract correspondences between computational models and human minds floats on very deep philosophical waters, and we will need to toss out a few lifelines before proceeding. I sketch a few ways to develop the "how actually" approach here, specifically drawing upon theories of explanation offered by Gualtiero Piccinini, Catherine Stinson, Lisa Miracchi, Rosa Cao, and Dan Yamins regarding the ways in which artificial neural network models might teach us something about the actual operations of the human mind that are crucial to its ability to create rational behavior, even if these models offer us only very partial pictures of its operations. Although these theories are closely related in their general take on the nature of explanation, I seek to highlight how each illuminates something distinctive that a DNN model might teach us about the organization of the human mind.[32]

Piccinini, for example, offers what he calls a "neurocomputational" account of how models might explain mental abilities (Piccinini 2015, 2020). The lynchpin of his view is the idea that artificial neural

[32] DNNs may help us understand human cognition even if they do not provide "how actually" explanations of it. Proofs of concept—that a physical system can learn to do something using only a certain kind of input data, for instance—can help defeat generic arguments that a certain type of data is necessary for human-like competence (Warstadt and Bowman 2022). "How possibly" models may also be useful indirectly in a variety of other roles in the context of discovery (Gelfert 2016:79–97). Nevertheless, I focus on the "how actually" strategy here, because I want to make the strongest case that can be supported by recent developments. I thank Alfredo Vernazzini for pushing me on this point.

network models might identify aspects of biological neural mechanisms that, when arranged in the right way, produce certain aspects of human cognition. "Aspect" is here a technical term; an aspect of a system is a property it possesses that is neither identical to nor entirely distinct from its lower-level constitution base, yet is synchronically determined by the arrangement of those lower-level constituents. Crucially, the causal properties possessed by the aspect of the system are some subset of its total causal properties. An artificial neural network system which recapitulates just the organizational structure that is responsible for producing those aspects in the target system might then be used to draw inferences about the target system, even if it does not implement its full suite of causal properties. In this sense, an artificial neural network can be treated as an empirical hypothesis concerning which abstract organizational features enable certain computational operations in the brain, such as an ability to efficiently extract certain types of abstract structure from certain types of data. Piccinini's neurocomputational account requires more than a proof of concept—in mechanistic terms, it privileges "how actually" explanation rather than merely "how possibly" explanation (Piccinini and Craver 2011)—because the aspects of the model that allow it to solve the task must correspond to real and relevant aspects of the human brain.

Catherine Stinson offers a similar view, which she calls a "generic mechanism" account of the explanatory relevance of artificial neural network models (Stinson 2020). Stinson's view is derived from numerous self-interpretive comments offered by prominent neural network modelers over the years, especially those which emphasize their desire to produce models which are "biologically plausible" (Stinson 2018). Stinson also seeks a conceptual framework that justifies drawing inferences from artificial models to biological target systems—in this case, reasoning that because something is true of a DNN, it should also be true of the human mind or brain. On Stinson's view, this inference cannot be justified on the basis of the model mimicking some behavior of the target system, for the two systems might achieve those outcomes in different ways. Moreover, we want a principle that justifies inferences about cases where the behaviors of the two systems are not already known to correspond, so that models can be used to teach us things that we do not already know about intelligence, rationality, or cognition. Stinson argues that the right relationship here is not one of mere behavioral mimicry, but rather shared membership in an abstract kind of mechanistic

structure, which she calls a "generic mechanism." In other words, we need to identify an abstract kind of mechanistic structure—in terms of types of structural properties like component types and connections amongst them—that could be shared between the brain and an DNN, and in virtue of which they exhibit some range of similar behaviors (Fig. 1.2).[33] This method places a bit more burden on the modeler, in that they must identify the kind of generic mechanism which is instantiated in both cases despite some prominent structural dissimilarities and (often) forms of idealization; but we will consider some generic mechanism candidates later in the book.

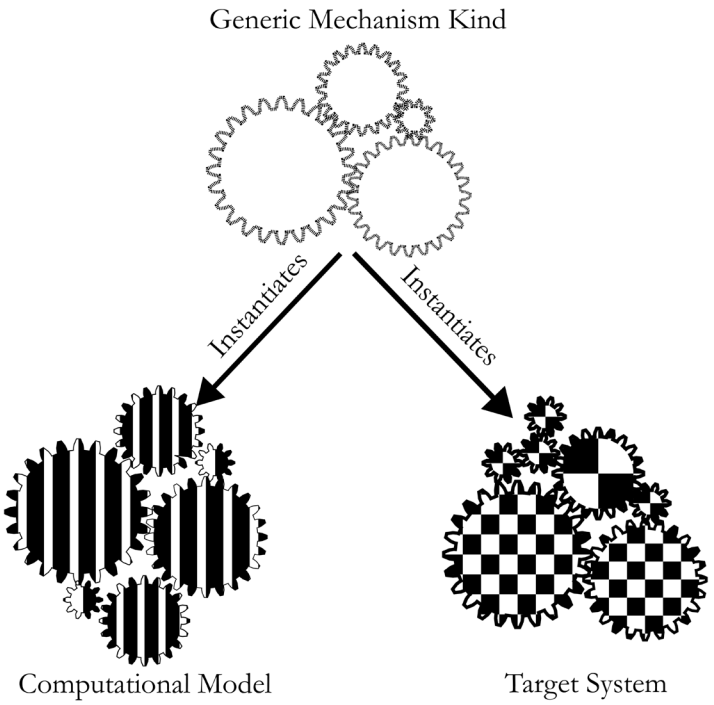


Figure 1.2 Stinson's "generic mechanism" view holds that a computational model (such as a DNN) can help explain phenomena implemented by a target system (such as the brain) if they both instantiate the same abstract mechanistic kind. Because they are both instances of the same generic mechanism, inferences about the computational model can generalize to the target system in virtue of their abstract mechanistic similarity. Of course, theorists should take care that the inferences generalized from the computational model to the target system issue from the abstract mechanistic structure they share, rather than from the specific ways in which they differ. (Figure adapted from Stinson 2020).

[33] In invoking the language of natural kinds here, Stinson perhaps goes a bit beyond Piccinini, in that the kinds must display stable clusters of inductive regularities that justify their inclusion in scientific practice (Boyd 1991, 1999; Khalidi 2013). However, there may be some persistent difficulties in identifying natural kinds of mechanism (Craver 2009), especially those at just the right level of abstraction (Boone and Piccinini 2016; Craver and Kaplan 2020). An alternative approach might be to use some of the tools of transformation-based abstraction taken from neuroscience and developed here later in Chapter 3 to establish a notion of "transform similarity" between the computational system and the target system to legitimize the abstract comparisons (Cao and Yamins 2021a, 2021b).

Finally, Lisa Miracchi (2019) holds an even more expansive view of the explanatory relevance of computational models. In fact, even those who believe that genuinely mental properties or distinctively human intelligence requires processing that is fundamentally different in kind from that which could occur in these computational models might be persuaded of deep learning's relevance to the human mind by considering Miracchi's view. Miracchi is concerned with "juicier" philosophical takes on mental properties, such as that all acts of mentality are essentially conscious or normative, that perception involves continuous reciprocal "enactive" engagement with the world, and/or that that consciousness or normativity may exhibit properties which cannot in principle be implemented in a machine. Someone who held such a view might deny that deep learning models share even aspects or generic mechanisms with the human mind, because consciousness, meaning, enactive engagement, or normativity are essential to even the mind's most humble acts of perception or inference. Miracchi thinks that even someone holding such a view should attend to developments in deep learning, because artificial neural network models might nevertheless teach us things about "generative difference-makers" in the human brain that are physical preconditions for human persons to engage in these minded or meaningful activities of these kinds (and see also Klein, Hohwy, and Bayne 2020; Miracchi 2017).

| Type of Model | Role in Explanation |
|---|---|
| Agent Model | "A model of the intelligence-related explanandum [agency, intelligence, perception, thinking, knowledge, language, rationality] that facilitates prediction and measurement" |
| Basis Model | "A model of the artificial system in computational, physical, and/or non-mental behavioral terms that facilitates manipulation and measurement" |
| Generative Model | "A model of how the features represented by the basis model make generative differences to features represented by the agent model." |

Table 1.2 The three types of models (together with their respective roles) recommended by Miracchi's (2019) generative difference-maker approach to explanation.

Miracchi's view is perhaps the most philosophically demanding of the three considered so far; her approach requires theorists to develop three separate background models which are connected to one another before we can decide how an artificial neural network model bears on the human mind. Namely, one needs an "agent model," which provides a theory of the relevant mental phenomenon (which includes perhaps ineliminable appeals to consciousness, normativity, or other intentional properties), a "basis model," which characterizes the artificial system in computational, mechanical, and/or behavioral terms, and a "generative" model, which explains how changes in the basis features make a difference to or determine features in the agent (Table 1.2). This third model might be seen to address one of the primary concerns about the "black box" nature of deep learning models: that while they often reproduce human or animal behaviors to a large degree, the lack of known "empirical links" tying the processing in the model to the processing of the human mind prevents us from gaining any new understanding of cognition or intelligence from the model's successes (Sullivan 2022). It is worth comparing this picture to the three-model approach of Cao and Yamins (2021a), who recommend developing a model of the target system, a model of the computational system, and a "transform similarity" mapping that shows how to map coarse-grained features from the computational and target models to one another in order to make predictions about how the effects of changes to the abstract dynamics in one system predicts effects of the abstract dynamics of the other (Fig. 1.3).

This book offers a response and illustration of this kind of interdisciplinary approach by drawing together many threads from thinkers in different disciplines. The philosophical and psychological research reviewed in the later chapters of this book can be interpreted as outlining agent models for mental faculties, and the computational research reviewed as providing basis models for targets which computer scientists have suggested might be relevant to understanding these mental properties. The third type of models (whether "generative" or "transform similarity") linking the other two remain to a large part unknown, but later chapters in this book can be seen to tee up their development for research groups up to the interdisciplinary challenge.

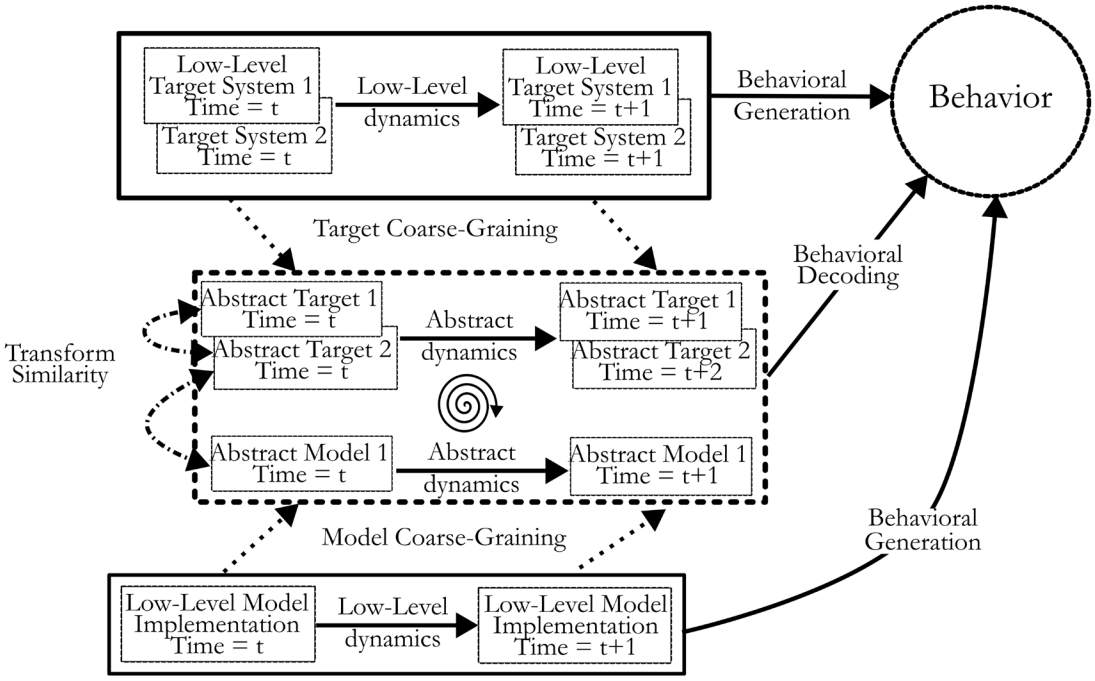


Figure 1.3 Cao & Yamins' (2022) "3M++" view on the explanatory relationship between computational models and underlying systems. Like Miracchi's view, their approach requires a third theory which specifies a "transform similarity" that allows one to map behavior-producing state transitions in a computational model to behavior-producing state transitions in a target system. They note that the low-level implementation details between one instance of a biological behavior-producing target system (e.g., a human) and another (e.g., another human or a member of a different species like a monkey) will also differ in their lowest-level organization (e.g., specific brain activity or neural state transitions) and need to be mappable into one another using the same method of transform similarity. (Figure adapted from Cao & Yamins 2022.)

### *1.6 Other dimensions of the rationalist-empiricist debate*

I argued above that the primary point of dispute between empiricism and nativism concerns the doctrine of innate ideas and the Transformation Principle. Many rationalists are nativists and many empiricists are anti-nativists, but other disagreements have accrued between these competing alliances over the last two millennia. Let us briefly consider and situate four other dimensions. First, there is a debate over the right way to *evaluate* the mental processes behind behavior, especially in terms of rationality (Taylor 2003). Rationalists tend to be "internalists"—that is, they think that mental processes are correct if they follow a set of rules, such as logic or decision theory. Empiricists, on the other hand, tend to be "externalists," evaluating

mental processes in terms of their reliability or success in particular environments. Second, there is a debate over the "format" of thought; rationalists tend to believe that thought occurs in a symbolic or language-like medium, often called a "Language of Thought" (Fodor 1975), which might be compared to symbolic computer programming code. Empiricists, on the other hand, tend to think that even abstract reasoning operates over sensorial, imagistic representations derived from perceptions (Barsalou 1999). These imagistic representations can be abstracted and recombined in various ways; but according to empiricists, they never fully lose their experiential content. Third, there is a debate over the justification of beliefs—with the "empiricist" side demanding that all empirical knowledge must be justified by objectively observable perceptual evidence (BonJour and Sosa 2003). Some flavor of this empiricist doctrine led psychological behaviorists like John Watson to banish all appeals to inner mental entities in psychology, because such theoretical entities cannot be objectively observed in experiments (Greenwood 1999). Fourth, there is a debate over the semantic content of meaningful claims, with "logical empiricists" insisting that all meaningful claims either be true in virtue of meaning alone (e.g. "all bachelors are unmarried males") or be translatable without remainder into "observation sentences" that should be true if the claim is correct (Quine 1951). This "semantic" strain of empiricism motivated the logical positivists in the Vienna Circle to dismiss most metaphysical questions as meaningless—such as whether the soul is immortal, or whether all events must have a cause—since these questions could not be translated into empirically testable hypotheses.

In this book, we will explore the first two additional dimensions of the rationalist-empiricist dispute to some degree but dismiss the latter two debates entirely. I will thus not attempt to revive the empiricisms of Skinner or Carnap here; I applaud Quine's (1951) exorcism of these previous "two dogmas of empiricism" in the middle of the last century. As noted above, we will broach the debate over the right way to evaluate cognition, because it is relevant to the problem of gauging progress towards human-like processing in AI. The dispute over the format of thought will also arise in a number of places throughout the book. This dimension of the debate will be finessed here because, while I agree that representations and representational structures must be extracted from sensory experience, they might be so heavily transformed by processing that by the

latest stages of processing, the signals might be indistinguishable from the sort of amodal symbols favored by critics of sensorial theories of concepts (Machery 2006).

*1.7 The DoGMA in relation to other recent revivals of empiricism*

Several other books have recently defended historical empiricism in the context of current cognitive science and with respect to neural network modeling specifically. In particular, Jesse Prinz (2002, 2004, 2007) revived and charitably applied the views of historical empiricists in contemporary cognitive science, and in many ways the DoGMA is continuous with this mission. Prinz, however, focused mostly on evidence for Lockean and Humean theses arising from evidence in the discipline of psychology, and he dealt more centrally with the debate over the format of thought (though he also endorsed anti-nativism about ideas—2005, p.679). The present book, however, is the first to update this empiricist frame in light of recent achievements in deep learning. In psychology and computational neuroscience, Jeffrey Elman and co-authors also challenged nativism using evidence from neural network modeling, in particular by emphasizing the promise of domain-general inductive biases (Elman et al. 1998). This book is in many ways a continuation of that approach for the deep learning era, whilst also pointing out how far the neural network modeling has come over just the last five years in vindicating even more ambitious empiricist proposals. Bill Bechtel and Adele Abrahamsen (Bechtel and Abrahamsen 2002) have also written on the relevance of neural network modeling to empiricism, though it is not their primary focus and they approach the question from a more philosophy of science perspective. Patricia Churchland (1989) and Paul Churchland (2012) have also championed neural network modeling against various versions of nativism and rationalism, though they tend to adopt a more eliminativist attitude towards the traditional mental categories invoked by historical empiricists and highlighted by the DoGMA, such as memory, imagination, and attention. Considering the DoGMA in light of these other recent defenses of empiricism will help put it in more complete historical and philosophical context.

*1.8 Basic strategy of the book: Understanding deep learning through empiricist faculty psychology*

Other excellent books are available which analyze the technical structure of deep learning architectures using the methodologies of computer science (e.g. Chollet 2021; e.g. Goodfellow, Bengio, and Courville 2016;

Zhang et al. 2021). The goal of this book is different: to provide an interdisciplinary interpretation of the previous accomplishments of these networks, a fair appraisal of what remains to be achieved, and an opinionated take on where they might profitably head next. Do deep learning's successes show that DNNs can model rational decision-making? Do their shortcomings show that they cannot? Can some architectures model some forms of rationality, but not others? In exploring these questions, I aim to give the reader not just an understanding of the technical minutiae or architectural choices which go into the design of a state-of-the-art DNN, but also an interdisciplinary, philosophically grounded account of the mental faculties that these networks can or cannot be fairly said to model.

The general idea that AI should focus on a modular cognitive architecture is not novel. In fact, it has been pursued by several different influential research groups over the years, especially those behind ACT-R, SOAR, Sigma, and CMC (Anderson and Lebiere 2014; Laird 2019; Laird, Lebiere, and Rosenbloom 2017; Rosenbloom 2013). The empiricist DoGMA proposed here, however, differs from these other approaches in several respects. For one, these approaches are usually hybrids of classical and connectionist approaches, often including both connectionist perceptual modules and classical rules or theories; the DoGMA proposed here aims for all modules to be network-based and to eschew innate domain-specific programming. For another, and to their credit, most of these other modular architectures have completed implementations that can be readily applied to data. The aim of this book is rather to provide philosophical and theoretical inspiration for the completion and full implementation of the DoGMA, which to date remains somewhat piecemeal in execution.

To be clear, none of these proposed cognitive architectures for AI, including the one suggested here, should be expected to be the complete and final draft of the mind's structure. The empiricists themselves differed in the details of their faculty taxonomies and theories of the individual faculties, and we should expect that empiricist modelers today will explore different possibilities as well. Here, I instead emphasize the domain-general modular strategy as the best way to understand and pursue the empiricist side of AI's current methodological debate.

### *1.9 Organization of the remaining chapters: Faculties, philosophers, and modules*

The next chapter provides an overview of deep learning and its characteristic strengths and weaknesses. The remaining chapters of the book offer an empiricist account of a mental faculty and explore attempts to integrate aspects of the faculty into a deep learning architecture. In the course of our exploration, we will explain how each faculty's activities might improve the performance of an artificial agent. We will ask whether the addition of the faculty enables the resulting architecture to achieve new kinds of rational decision-making. Chapter 2 lays out some methodological preliminaries of this task, including specific tiers of rationality and principles for fair evaluation. Chapter 3 focuses on the faculty of perception and how DCNNs might reveal generative difference-makers in human perception by deploying structures that perform multiple forms of abstraction from sensory experience emphasized by historical empiricists. Chapter 4 reviews attempts to model the role of memory in deep learning systems, focusing on architectures like the Episodic Controller (EC). Chapter 5 discusses attempts to model the imagination using architectures like GANs and Variational Autoencoders (VAEs), and Chapter 6 highlights various tweaks inspired by aspects of attention, such as transformers. Chapter 7 reviews more nascent attempts to model human social and moral cognition—focusing especially on the faculty of empathy/sympathy—ending with some speculative suggestions for future research to model cultural and social learning in empiricist, DNN-based agents.

The course of each chapter reflects the book's interdisciplinary goals. Because I aim to illustrate the value of revisiting historical empiricism in arbitrating current debates over deep learning, I review the theories offered by empiricist philosophers for these psychological faculties. While I could easily survey the views of a dozen empiricists on each faculty, for the sake of narrative simplicity each faculty chapter zooms in on the views of one prominent empiricist-leaning philosopher who had a particularly rich and germane take on the faculty highlighted in that chapter. These philosophers include John Locke, Ibn Sina, David Hume, William James, and Sophie de Grouchy. This "philosopher of the week" structure is deployed as a narrative device, so that we might go into enough depth with the views of a wide sample of empiricist philosophers. This illustrates the utility of empiricist philosophy generally as a source for engineering inspiration, while allowing us to avoid entrapment in the details of particular interpretive disputes. After sketching the philosophical and psychological landscape, each chapter then reviews current and future implementations of the processing

attributed to this faculty in DNN architectures, often on the basis of explicit comparisons by the DNN modelers. We will also routinely review the degree of biological plausibility of the efforts thus far, by discussing similarities between the architecture's structure and the neural circuits hypothesized to implement the faculty in biological brains.

While I argue that the DoGMA already inspires and is in turn bolstered by a variety of recent achievements in deep learning, the models behind these recent achievements mostly implement only one or two additional faculty modules, and new problems will arise when we start trying to coordinate more complex DNN-based architectures featuring multiple semi-independent faculties. Here, we will study these issues under the heading of the Control Problem. No existing DNN-based system attempts to model all of the faculties explored here; most contain modules for one or more forms of memory and some model attentional resources, but few have explored imagination or social cognition and none simultaneously include components for all the faculties canvassed here.[34] This has sometimes been a goal of past AI projects—some implementations of ACT-R, for example, included components corresponding to many of the faculties discussed here. We may expect that problems of coordination and control will become more pressing as more faculties are included and agent architectures become more complex. Evolution balanced these control problems through iterative parameter search during millions of years of natural selection, and we may expect a long period of exploration will be required in AI as well. I hope to inspire AI researchers to consider these control problems more proactively, and to begin working on the design and implementation of more ambitious faculty architectures that model the integrated decision-making of a whole rational agent, rather than focusing on one or two components at a time. We will return to re-evaluate this challenge and call to action in the final chapter before closing. Thus, to work.

[34] This is generally acknowledged by the authors of these alternative architectures. For example, Laird, Lebiere, and Rosenbloom recently wrote that their model "remains incomplete in a number of ways…for example, concerning metacognition, emotion, mental imagery, direct communication and learning across modules, the distinction between semantic and episodic memory, and mechanisms necessary for social cognition" (2017, p23).

**References**

Achille, Alessandro, and Stefano Soatto. 2018. "Emergence of Invariance and Disentanglement in Deep Representations." *Journal of Machine Learning Research* 19(50):1–34.

Anderson, John R., and Christian J. Lebiere. 2014. *The Atomic Components of Thought*. Psychology Press.

Ariew, Andre. 1996. "Innateness and Canalization." *Philosophy of Science* 63:S19–27.

Baldi, Pierre, Peter Sadowski, and Daniel Whiteson. 2014. "Searching for Exotic Particles in High-Energy Physics with Deep Learning." *Nature Communications* 5(1):1–9.

Barsalou, Lawrence W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22(4):577–660.

Bechtel, William, and Adele Abrahamsen. 2002. *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*. Blackwell Publishing.

Berwick, Robert C., and Noam Chomsky. 2016. *Why Only Us: Language and Evolution*. MIT press.

Berwick, Robert C., and Noam Chomsky. 2017. "Why Only Us: Recent Questions and Answers." *Journal of Neurolinguistics* 43:166–77. doi: 10.1016/j.jneuroling.2016.12.002.

BonJour, Laurence, and Ernest Sosa. 2003. *Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues*. Oxford: Wiley-Blackwell.

Boone, Worth, and Gualtiero Piccinini. 2016. "Mechanistic Abstraction." *Philosophy of Science* 83(5):686–97.

Botvinick, Matthew, David GT Barrett, Peter Battaglia, Nando de Freitas, Darshan Kumaran, Joel Z. Leibo, Timothy Lillicrap, Joseph Modayil, Shakir Mohamed, and Neil C. Rabinowitz. 2017. "Building Machines That Learn and Think for Themselves." *Behavioral and Brain Sciences* 40.

Boyd, Richard. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61(1):127–48.

Boyd, Richard. 1999. "Kinds, Complexity and Multiple Realization." *Philosophical Studies* 95(1):67–98.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models Are Few-Shot Learners." Pp. 1877–1901 in *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc.

Bruna, Joan, and Stéphane Mallat. 2011. "Classification with Scattering Operators." Pp. 1561–66 in *CVPR 2011*. IEEE.

Buckner, Cameron. 2013. "Morgan's Canon, Meet Hume's Dictum: Avoiding Anthropofabulation in Cross-Species Comparisons." *Biology & Philosophy* 28(5):853–71.

Buckner, Cameron. 2018. "Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks." *Synthese* 195(12):5339–72.

Buckner, Cameron. 2022. "A Forward-Looking Theory of Content." *Ergo* 8(37):367–401.

Campbell, Murray. 1999. "Knowledge Discovery in Deep Blue." *Communications of the ACM* 42(11):65–67. doi: 10.1145/319382.319396.

Cao, Rosa, and Daniel Yamins. 2021a. "Explanatory Models in Neuroscience: Part 1–Taking Mechanistic Abstraction Seriously." *ArXiv Preprint ArXiv:2104.01490*.

Cao, Rosa, and Daniel Yamins. 2021b. "Explanatory Models in Neuroscience: Part 2–Constraint-Based Intelligibility." *ArXiv Preprint ArXiv:2104.01489*.

Carey, Susan, and Elizabeth Spelke. 1996. "Science and Core Knowledge." *Philosophy of Science* 63(4):515–33.

Carruthers, Peter. 2006. *The Architecture of the Mind*. Oxford University Press.

Childers, Timothy, Juraj Hvorecký, and Ondrej Majer. 2021. "Empiricism in the Foundations of Cognition." *AI & SOCIETY*. doi: 10.1007/s00146-021-01287-w.

Chollet, François. 2019. "On the Measure of Intelligence." *ArXiv Preprint ArXiv:1911.01547*.

Chollet, Francois. 2021. *Deep Learning with Python*. Simon and Schuster.

Chomsky, Noam. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Cambridge University Press.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Greenwood Publishing Group.

Chomsky, Noam. 1993. "A Minimalist Program for Linguistic Theory." in *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*, edited by K. Hale and S. J. Keyser. Cambridge, MA: MIT Press.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann. 2022. "Palm: Scaling Language Modeling with Pathways." *ArXiv Preprint ArXiv:2204.02311*.

Christiansen, Morten H., and Nick Chater. 2016. *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA, USA: MIT Press.

Churchland, Patricia Smith. 1989. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. MIT press.

Churchland, Paul M. 2012. *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. MIT press.

Conze, Edward. 1963. "Spurious Parallels to Buddhist Philosophy." *Philosophy East and West* 13(2):105–15.

Cowie, Fiona. 1998. "Mad Dog Nativism." *The British Journal for the Philosophy of Science* 49(2):227–52.

Craver, C. F. 2009. "Mechanisms and Natural Kinds." *Philosophical Psychology* 22(5):575–94. doi: 10.1080/09515080903238930.

Craver, Carl F., and David M. Kaplan. 2020. "Are More Details Better? On the Norms of Completeness for Mechanistic Explanations." *The British Journal for the Philosophy of Science* 71(1):287–319.

Dąbrowska, Ewa. 2015. "What Exactly Is Universal Grammar, and Has Anyone Seen It?" *Frontiers in Psychology* 6:852.

De Brigard, Felipe. 2014. "Is Memory for Remembering? Recollection as a Form of Episodic Hypothetical Thinking." *Synthese* 191(2):155–85.

Demeter, Tamás. 2021. "Fodor's Guide to the Humean Mind." *Synthese* 199(1):5355–75.

Elman, Jeffrey L., Elizabeth A. Bates, and Mark H. Johnson. 1998. *Rethinking Innateness: A Connectionist Perspective on Development*. Vol. 10. MIT press.

Evans, Richard. 2020. "Kant's Cognitive Architecture." PhD Thesis, Imperial College London.

Ferrucci, David, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, and John Prager. 2010. "Building Watson: An Overview of the DeepQA Project." *AI Magazine* 31(3):59–79.

Fodor, Janet Dean. 2001. "Setting Syntactic Parameters." in *The handbook of Contemporary Syntactic Theory*, edited by M. Baltin and C. Collins. New York: Blackwell.

Fodor, Jerry A. 1975. *The Language of Thought*. Harvard University Press.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT press.

Fodor, Jerry A. 2008. *LOT 2: The Language of Thought Revisited*. Oxford University Press on Demand.

Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28(1–2):3–71.

Frank, Michael C., Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, Erin E. Hannon, Melissa Kline, Claartje Levelt, Casey Lew-Williams, Thierry Nazzi, Robin Panneton, Hugh Rabagliati, Melanie Soderstrom, Jessica Sullivan, Sandra Waxman, and Daniel Yurovsky. 2017. "A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building." *Infancy* 22(4):421–35. doi: 10.1111/infa.12182.

Fukushima, Kunihiko, and Sei Miyake. 1982. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition." Pp. 267–85 in *Competition and cooperation in neural nets*. Springer.

Gauker, Christopher. 2013. *Words and Images: An Essay on the Origin of Ideas*. Reprint edition. Oxford: Oxford University Press.

Gelfert, Axel. 2016. *How to Do Science with Models: A Philosophical Primer*. Springer.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

Gopnik, Alison. 2009. "Could David Hume Have Known about Buddhism?: Charles François Dolu, the Royal College of La Flèche, and the Global Jesuit Intellectual Network." *Hume Studies* 35(1/2):5–28.

Goyal, Anirudh, and Yoshua Bengio. 2020. "Inductive Biases for Deep Learning of Higher-Level Cognition." *ArXiv Preprint ArXiv:2011.15091*.

Greenwood, John D. 1999. "Understanding the'Cognitive Revolution' in Psychology." *Journal of the History of the Behavioral Sciences* 35(1):1–22.

Griffiths, Paul E., and Edouard Machery. 2008. "Innateness, Canalization, and 'Biologicizing the Mind.'" *Philosophical Psychology* 21(3):397–414.

Griffiths, Paul E., and James Tabery. 2013. "Developmental Systems Theory: What Does It Explain, and How Does It Explain It?" Pp. 65–94 in *Advances in child development and behavior*. Vol. 44. Elsevier.

Hasson, Uri, Samuel A. Nastase, and Ariel Goldstein. 2020. "Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks." *Neuron* 105(3):416–34.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hubel, David H., and Torsten N. Wiesel. 1967. "Cortical and Callosal Connections Concerned with the Vertical Meridian of Visual Fields in the Cat." *Journal of Neurophysiology* 30(6):1561–73.

Janiak, Andrew. 2020. *Space: A History*. New York: Oxford University Press.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and Anna Potapenko. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596(7873):583–89.

Kampis, Dora, Petra Karman, Gergely Csibra, Victoria Southgate, and Mikolaj Hernik. 2020. "A Two-Lab Direct Replication Attempt of Southgate, Senju, & Csibra (2007)."

Khalidi, Muhammad Ali. 2001. "Innateness and Domain Specificity." *Philosophical Studies* 105(2):191–210.

Khalidi, Muhammad Ali. 2013. *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*. Cambridge University Press.

Khalidi, Muhammad Ali. 2016. "Innateness as a Natural Cognitive Kind." *Philosophical Psychology* 29(3):319–33.

Klein, Colin, Jakob Hohwy, and Tim Bayne. 2020. "Explanation in the Science of Consciousness: From the Neural Correlates of Consciousness (NCCs) to the Difference Makers of Consciousness (DMCs)." *Philosophy and the Mind Sciences* 1(II).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." Pp. 1097–1105 in *Advances in neural information processing systems*.

Kulke, Louisa, Britta von Duhn, Dana Schneider, and Hannes Rakoczy. 2018. "Is Implicit Theory of Mind a Real and Robust Phenomenon? Results from a Systematic Replication Study." *Psychological Science* 29(6):888–900.

Laird, John E. 2019. *The Soar Cognitive Architecture*. MIT press.

Laird, John E., Christian Lebiere, and Paul S. Rosenbloom. 2017. "A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics." *Ai Magazine* 38(4):13–26.

Landy, David. 2017. *Hume's Science of Human Nature: Scientific Realism, Reason, and Substantial Explanation*. Routledge.

Lasnik, Howard, and Terje Lohndal. 2010. "Government–Binding/Principles and Parameters Theory." *Wiley Interdisciplinary Reviews: Cognitive Science* 1(1):40–50.

Laurence, Stephen, and Eric Margolis. 2002. "Radical Concept Nativism." *Cognition* 86(1):25–55.

Laurence, Stephen, and Eric Margolis. 2012. "Abstraction and the Origin of General Ideas." *Philosopher's Imprint* 12(19).

Laurence, Stephen, and Eric Margolis. 2015. "Concept Nativism and Neural Plasticity." Pp. 117–47 in *The conceptual mind: New directions in the study of concepts*, edited by S. Laurence and E. Margolis. MIT Press.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1(4):541–51.

Lorenz, Konrad. 1935. "Der Kumpan in Der Umwelt Des Vogels. Der Artgenosse Als Auslösendes Moment Sozialer Verhaltungsweisen." *Journal Für Ornithologie. Beiblatt.(Leipzig)*.

Machery, Edouard. 2006. "Two Dogmas of Neo-Empiricism." *Philosophy Compass* 1(4):398–412.

Machery, Edouard. 2009. *Doing without Concepts*. Oxford University Press.

Mallon, Ron, and Jonathan M. Weinberg. 2006. "Innateness as Closed Process Invariance." *Philosophy of Science* 73(3):323–44.

Mameli, Matteo, and Patrick Bateson. 2006. "Innateness and the Sciences." *Biology and Philosophy* 21(2):155–88.

Marcus, Gary. 2018. "Innateness, Alphazero, and Artificial Intelligence." *ArXiv Preprint ArXiv:1801.05667*.

Matuszek, Cynthia, Michael Witbrock, John Cabral, and John DeOliveira. 2006. "An Introduction to the Syntax and Content of Cyc." *UMBC Computer Science and Electrical Engineering Department Collection*.

McClelland, James L., Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. 2010. "Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition." *Trends in Cognitive Sciences* 14(8):348–56.

Millican, Peter. 2009. "Hume on Induction and the Faculties." *Draft Article*.

Miracchi, Lisa. 2017. "Generative Explanation in Cognitive Science and the Hard Problem of Consciousness." *Philosophical Perspectives* 31(1):267–91.

Miracchi, Lisa. 2019. "A Competence Framework for Artificial Intelligence Research." *Philosophical Psychology*.

Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. Penguin UK.

Mole, Christopher. 2011. *Attention Is Cognitive Unison: An Essay in Philosophical Psychology*. Oxford University Press.

Montalvo, David. 1999. "The Buddhist Empiricism Thesis: An Extensive Critique." *Asian Philosophy* 9(1):51–70.

von Neumann, J. 1993. "First Draft of a Report on the EDVAC." *IEEE Annals of the History of Computing* 15(4):27–75. doi: 10.1109/85.238389.

Newell, Allen, and Herbert A. Simon. 1976. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM* 19(3):113–26.

Northcott, Robert, and Gualtiero Piccinini. 2018. "Conceived This Way: Innateness Defended." *Philosophers' Imprint* 18(18).

Oakes, Lisa M. 2017. "Sample Size, Statistical Power, and False Conclusions in Infant Looking-Time Research." *Infancy: The Official Journal of the International Society on Infant Studies* 22(4):436–69. doi: 10.1111/infa.12186.

Olsson, Catherine, Elhage Nelson, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandish, and Chris Olah. 2022. "In-Context Learning and Induction Heads." *Transformer Circuits*. Retrieved (https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html).

Pearl, Judea. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc.

Pearl, Judea. 2019. "The Seven Tools of Causal Inference, with Reflections on Machine Learning." *Communications of the ACM* 62(3):54–60.

Pearl, Judea. 2021. "Radical Empiricism and Machine Learning Research." *Journal of Causal Inference* 9(1):78–82.

Pessoa, Luiz, Loreta Medina, and Ester Desfilis. 2022. "Refocusing Neuroscience: Moving Away from Mental Categories and towards Complex Behaviours." *Philosophical Transactions of the Royal Society B* 377(1844):20200534.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanistic Account*. OUP Oxford.

Piccinini, Gualtiero. 2020. *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford University Press.

Piccinini, Gualtiero, and Carl Craver. 2011. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183(3):283–311. doi: 10.1007/s11229-011-9898-4.

Pinker, Steven. 2003. *The Blank Slate: The Modern Denial of Human Nature*. Penguin.

Powers, John. 1994. "Empiricism and Pragmatism in the Thought of Dharmakīrti and William James." *American Journal of Theology & Philosophy* 15(1):59–85.

Prinz, Jesse. 2007. *The Emotional Construction of Morals*. Oxford University Press.

Prinz, Jesse J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press.

Prinz, Jesse J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press.

Quine, W. V. 1951. "Two Dogmas of Empiricism." *The Philosophical Review* 60(1):20–43.

Quine, Willard Van Orman. 1969. "Linguistics and Philosophy." in *Language and Philosophy*, edited by S. Hook. New York University Press.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. "Hierarchical Text-Conditional Image Generation with Clip Latents." *ArXiv Preprint ArXiv:2204.06125*.

Ritchie, J. Brendan. 2021. "What's Wrong with the Minimal Conception of Innateness in Cognitive Science?" *Synthese* 199(1):159–76.

Robbins, Philip. 2009. "Modularity of Mind" edited by E. N. Zalta. *The Stanford Encyclopedia of Philosophy*.

Rogers, Timothy T., and James L. McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT press.

Ronacher, Bernhard. 2019. "Innate Releasing Mechanisms and Fixed Action Patterns: Basic Ethological Concepts as Drivers for Neuroethological Studies on Acoustic Communication in Orthoptera." *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology* 205(1):33–50. doi: 10.1007/s00359-018-01311-3.

Rosenblatt, Frank. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65(6):386.

Rosenbloom, Paul S. 2013. "The Sigma Cognitive Architecture and System." *AISB Quarterly* 136:4–13.

Rubio-Fernández, Paula. 2019. "Publication Standards in Infancy Research: Three Ways to Make Violation-of-Expectation Studies More Reliable." *Infant Behavior and Development* 54:177–88.

Salvadori, Eliala, Tatiana Blazsekova, Agnes Volein, Zsuzsanna Karap, Denis Tatone, Olivier Mascaro, and Gergely Csibra. 2015. "Probing the Strength of Infants' Preference for Helpers over Hinderers: Two Replication Attempts of Hamlin and Wynn (2011)." *PLOS ONE* 10(11):e0140570. doi: 10.1371/journal.pone.0140570.

Samuels, Richard. 2004. "Innateness in Cognitive Science." *Trends in Cognitive Sciences* 8(3):136–41.

Samuels, Richard. 2007. "Is Innateness a Confused Notion?" in *The Innate Mind: Foundations and the Future*, edited by P. Carruthers, S. Laurence, and S. Stich. Oxford University Press.

Shallue, Christopher J., and Andrew Vanderburg. 2018. "Identifying Exoplanets with Deep Learning: A Five-Planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90." *The Astronomical Journal* 155:94.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550(7676):354–59.

Smith, Linda B. 2000. "Avoiding Associations with It's Behaviorism You Really Hate." Pp. 169–74 in *Becoming a word learner A debate on lexical acquisition*, edited by R. Golinkoff. Oxford University Press.

Spelke, Elizabeth. 1994. "Initial Knowledge: Six Suggestions." *Cognition* 50(1–3):431–45.

Sterelny, Kim. 1989. "Fodor's Nativism." *Philosophical Studies* 55(2):119–41.

Stinson, Catherine. 2018. "Explanation and Connectionist Models." Pp. 120–34 in *The Routledge Handbook of the Computational Mind*, edited by M. Sprevak and M. Colombo. Routledge.

Stinson, Catherine. 2020. "From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence." *Philosophy of Science* 87(4):590–611.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73(1):109–33. doi: 10.1093/bjps/axz035.

Sutton, Richard. 2019. "The Bitter Lesson." *Incomplete Ideas*. Retrieved November 3, 2020 (http://www.incompleteideas.net/IncIdeas/BitterLesson.html).

Taylor, Kenneth. 2003. *Reference and the Rational Mind*. CSLI Publications.

The ManyBabies Consortium. 2020. "Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference." *Advances in Methods and Practices in Psychological Science* 3(1):24–52. doi: 10.1177/2515245919900809.

Tillemans, Tom. 2021. "Dharmakīrti." in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. Metaphysics Research Lab, Stanford University.

Turing, Allen M. 1950. "Computing Machinery and Intelligence." *Mind* 59(236):433.

Uzgalis, William. 2020. "John Locke." in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. Metaphysics Research Lab, Stanford University.

Van Leeuwen, Neil. 2013. "The Meanings of 'Imagine' Part I: Constructive Imagination." *Philosophy Compass* 8(3):220–30.

Vinyals, Oriol, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech Czarnecki, Andrew Dudzik, Aja Huang, and ... 2019. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II." *DeepMind*. Retrieved April 5, 2019 (https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/).

Warstadt, Alex, and Samuel R. Bowman. 2022. "What Artificial Neural Networks Can Tell Us about Human Language Acquisition." Pp. 17–60 in *Algebraic Structures in Natural Language*. CRC Press.

Woodfield, Andrew. 1991. "Conceptions." *Mind* 100(4):547–72.

Zaadnoordijk, Lorijn, Tarek R. Besold, and Rhodri Cusack. 2022. "Lessons from Infant Learning for Unsupervised Machine Learning." *Nature Machine Intelligence* 4(6):510–20.

Zador, Anthony M. 2019. "A Critique of Pure Learning and What Artificial Neural Networks Can Learn from Animal Brains." *Nature Communications* 10(1):1–7. doi: 10.1038/s41467-019-11786-6.

Zhang, Aston, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 2021. "Dive into Deep Learning." *ArXiv Preprint ArXiv:2106.11342*.