Robert Long
rl2898@nyu.edu
Draft: 30 September 2020

Abstract: Recent progress in AI has revitalized the debate between nativists and empiricists. In this paper, I characterize the nativism/empiricism dispute in theories of human intelligence, broadening this account in order to apply it to artificial intelligence. Characterizing AI systems as "nativist" or "empiricist", I taxonomize varieties of AI empiricism and AI nativism. I then consider arguments for these empiricist and nativist positions in AI. We will see that there is a reasonably strong case that empiricist human-level AI is at the very least possible. That is, it is unlikely that nativist starting points are *necessary* for human-level AI, although there may be some role for nativist architectures at an intermediate stage. It remains an open question, however, whether nativist or empiricist approaches will be more successful in practice as a path to human-level AI. Furthermore, there may be strong normative reasons to build nativist AI.
Word count: 10,123

Nativism and empiricism in artificial intelligence

Introduction

The last ten years have witnessed astonishing progress in AI, driven in large part by advances in machine learning and increases in data and computational power.

For example, the systems AlphaGo, AlphaGo Zero, and AlphaZero have achieved progress that had seemed out of reach just a few years prior.[1] AlphaZero excites so many people because it learns to play Go, chess, and shogi "from scratch," in the sense that it lacks the kinds of game-specific, hand-crafted features that past game-playing systems like DeepBlue had used—such as heuristics for what a good position is, or databases of proven opening or closing strategies. AlphaZero's engineers claim that AlphaZero shows that "a general-purpose reinforcement learning algorithm can achieve, tabula rasa, superhuman performance across many challenging domains."[2] This is a deliberate allusion to the empiricism of Locke.

Indeed, key elements of the DeepMind research strategy are self-consciously *empiricist*: building systems with very little domain-specific knowledge, and letting them learn what they need from data using domain-general learning algorithms. The success of AlphaZero, and advances in

[1] "The Mystery of Go, the Ancient Game That Computers Still Can't Win | WIRED." 2014. Accessed February 20, 2019. https://www.wired.com/2014/05/the-world-of-computer-go/.
[2] Silver et al. (2018)

deep learning and reinforcement learning more generally, have given AI a relatively empiricist bent in the last decade, at least in how AI researchers conceive of what they are doing.

One can also see this empiricist bent in recent natural language processing. While some approaches to natural language processing emphasize the need for symbol manipulation and syntactic structures, drawing inspiration from how humans are thought to process language, OpenAI's GPT-3 eschews such techniques and instead leverages massive amounts of data and a conceptually simple transformer architecture. In doing so, GPT-3 achieves progress in text generation without leveraging the language-specific priors that one might have thought would be required for these tasks.[3]

Historically, the dispute between empiricists and nativists has been about the nature of human and animal minds. But empiricist and nativist concerns now arise in artificial intelligence. This paper uses nativism and empiricism to address questions about the nature of artificial intelligence.[4] Will continued progress towards human-level AI (HLAI) require the construction of nativist systems, or can human-level AI systems be wholly or largely empiricist?

In this paper, I first characterize the nativism/empiricism dispute in theories of human (and animal) intelligence. I then broaden this account in order to apply it to artificial intelligence, and to characterize AI systems as "nativist" or "empiricist." Next, I taxonomize varieties of AI empiricism and AI nativism. Finally, I consider arguments for these various empiricist and nativist positions in AI. We will see that there is a reasonably strong case that empiricist human-level AI is at the very least possible. That is, it is unlikely that nativist starting points are *necessary* for human-level AI, although there may be some role for nativist architectures at an intermediate stage. It remains an open question, however, whether nativist or empiricist approaches will be more successful in practice as a path to human-level AI. Furthermore, there may be strong normative reasons to build nativist AI.

I. Broadening nativism and empiricism for artificial intelligence

In cognitive science, nativists and empiricists seek to explain what the infant mind must be like in order for learning and development to be possible.

Consider language acquisition. Chomsky's theory of language acquisition is the *locus classicus* of modern nativism. Chomsky asked: given the data that children are exposed to—their

[3]Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. "Language Models Are Few-Shot Learners." ArXiv, arXiv–2005. Of course, GPT-3 is still far from human-level language production and seems to be doing something very different from what humans are doing (Lake and Murphy ms); how far we can expect these techniques to scale is an open question.

[4] For another kind of connection, see Buckner's (2018) fascinating examination of how convolutional neural networks help shed light on traditional questions about abstraction in empiricist theories of the mind.

"primary linguistic data"—what psychological traits must be present independent of these experiences, in order for language acquisition to occur? Chomsky argues that children are equipped with an innately specified Language Acquisition Device, which encodes knowledge of Universal Grammar, or the set of possible grammars of natural languages. Without this innate, domain-specific endowment, Chomsky argues, the primary linguistic data would not be enough for children to learn the grammar which they do, in fact, reliably learn. The primary linguistic data would, if not supplemented by domain-specific innate knowledge, underdetermine the correct grammar.[5]

Or consider our ability to track objects and to expect that they will behave in certain ways. Mature humans single out objects and expect them to persist through occlusion, block each other's trajectories, and move continuously. What accounts for this feature of the mind? According to the "core cognition" (or "core knowledge") framework of nativists like Susan Carey and Elizabeth Spelke, infants have an innate concept of "object" that is not itself learned from experience, but which enables subsequent learning from experience.[6] [7]

In contrast with these nativist theories, consider B. F. Skinner's behaviorism. Skinner does not posit innate traits or concepts specific to language learning (like Chomsky's Universal Grammar) or to object perception (like a core cognitive system for objects) or to intuitive physics. Instead, Skinner posits a single domain-general learning mechanism that underlies psychological development across all domains. The learning mechanism is the same for the acquisition of language, object perception, and intuitive physics: namely, creatures increase behaviors in response to positive reinforcement and decrease them in response to negative reinforcement.[8]

How should we understand the dispute between nativists and empiricists? What makes "nativism" and "empiricism" helpful ways of carving up opposing theories? Here's one common proposal: nativists think that the mind has innate, unlearned features, whereas empiricists think everything is learned. However, this cannot really be the dispute. As Samet (1987) points out, "Everyone agrees that learning requires that *something* be innate." [9]

Margolis and Laurence (2012) give an illuminating characterization of the disagreement, in terms of domain-specificity and domain-generality.[10] Drawing from their account, in this paper I

---

[5] Chomsky, Noam. Aspects of the theory of syntax. Cambridge, MA: MITPress. 1965.
[6] Carey, Susan. The Origin of Concepts. Oxford University Press; 2009. 609 p.
[7] Spelke, Elizabeth. Initial knowledge: six suggestions. Cognition. 1994 Apr 1;50(1):431–45.
[8] Skinner, BF. Verbal behavior. New York: Appleton-Century-Crofts; 1957. For more modern versions of empiricism, see (among many others) Prinz (2002), Barsalou (1999), Elman et al. (1998).
[9]Samet, Jerry. Troubles with Fodor's nativism. Midwest Studies in Philosophy. 1987;10(1):575–594. p575
[10]Margolis and Laurence write that nativism posits "many distinct types of mechanisms, states, and processes for acquiring psychological traits, and supposes that different systems of acquisition operate across different psychological domains," whereas empiricism posits "few distinct types of mechanisms, states, and processes for acquiring psychological traits, and supposes that the same systems of acquisition operate across many psychological domains." I find this

give the following gloss on nativist and empiricist systems *in general*, which can apply to both biological and artificial systems:

> A nativist system is one whose initial state contains domain-specific mechanisms, states, and processes.
>
> An empiricist system is one whose initial state contains only domain-general mechanisms, states, and processes.

What do we mean by domain-specific or domain-general? There is a domain of language acquisition, and a nativist approach to language acquisition says that there are innate language-specific mechanisms, states, or processes. Domains can also be quite broad: for example, Susan Carey hypothesizes an innate "object system," "agency system," and "number system." These are (according to Carey) separate domains, with relatively encapsulated systems of acquisition and their own principles governing, even if the concepts they deal with apply very widely.

Thus, we can say that the key issue of nativism and empiricism in cognitive science is whether humans (or non-human animals) come equipped with only domain-general mechanisms, states, or processes, or whether they have domain-specific mechanisms, states, or processes. That is, whether they are empiricist or nativist systems. The notions of empiricist and nativist system allow us to unify the discussions of empiricism and nativism in cognitive science and in AI. In cognitive science, the debate is about whether *actually existing systems* (humans and non-human animals) are nativist or empiricist systems. In AI, the debate is about what systems it is *possible* and perhaps *practical* to build. The issue is what kind of initial state of an artificial system is necessary, or practical, for that system to be able to learn from data and reach some desired target state with some capacities. For example, you might wonder what one would need, or want, to build into a system in order for it to be able to learn to play Civilization, or learn to tell jokes; these are questions about what's possible or practical for the reaching of some hypothetical target state. For example, DeepMind's AlphaZero paper above is making a claim about what it is possible for general-purpose reinforcement learning algorithms to achieve.

What is required for building AI systems with dedicated capacities like Go-playing or translation is a rich and important question. But for present purposes, the central issue we will consider is whether domain-general learning mechanisms can achieve human-level success across

---

definition to be extremely fruitful and clarifying for understanding the debate in cognitive science. For the purposes of this paper, I have dropped the "many" and "few" element. This means that an AI with many "distinct types of mechanisms," all of which are domain-general (for example domain-general reinforcement learning, convolutions, and Monte Carlo Tree Search), will count as an empiricist AI, not a nativist AI. In addition, my notion of empiricism adds a rather strong requirement: *only* domain-general mechanisms. This makes nativism and empiricism exclusive and exhaustive.

many domains in a flexible way. That is, the central issue concerns possible paths to human-level artificial general intelligence (HLAI). HLAI requires systems that achieve human-level or superhuman performance at a wide range of the tasks typically associated with intelligence: not just playing Go, but also linguistic competence, common sense, the ability to navigate an environment, and so forth. Human-level AI is indexed to *performance*. On this usage, an HLAI need not perform tasks in the same way that humans perform them. AlphaZero counts as a superhuman task AI even though it doesn't play or learn Go in a human-like way.[11]

This characterization helps illuminate different approaches to AI research. Empiricist AI approaches emphasize a single, domain-general, learning mechanism. Consider the following statements from AI practitioners. DeepMind (2017) proposes "inbuilt knowledge [that] takes a highly generic form, leaving wide scope for learning to absorb domain-specific structure . . . avoiding a dependence on detailed, domain-specific prior information." Yann LeCun (2017) claims that none of innate machinery proposed by AI nativist Gary Marcus (see below) is necessary for human-level AI.

These approaches are empiricist in that they seek to build *all* intelligence traits, across many domains, via general learning processes. LeCun is basically denying necessity nativism, at least of Marcus's sort. DeepMind seems to be advocating for the possibility (or practicality) of empiricism: we don't need inbuilt knowledge and mechanisms specific to language, or inbuilt knowledge and mechanisms specific to object perception. Instead, some minimal set of learning mechanisms can lead to intelligent behavior across different domains.

What about AI nativism? The most prominent exponent of AI nativism today is Gary Marcus. He thinks that different domains of learning will require human-level AI to have domain-specific innate "representation, algorithms, and knowledge." He explicitly ties his position to nativism about humans.

> Ultimately, it seems likely that many different types of tasks will have their own innate requirements: Monte Carlo tree search for board games, syntactic tree manipulation operations for language understanding, geometric primitives for 3-D scene understanding, theory of mind for problems demanding social coalitions, and so forth. Taken together, the full set of primitives may look less like a tabula rasa and more like . . . the sort of things that strong nativists like myself, Noam Chomsky, Elizabeth Spelke, Steve Pinker and the late Jerry Fodor have envisioned.

[11] Although there is some skepticism about the possibility of human-level artificial intelligence, or the coherence of the notion, in this paper I take for granted, as do many AI researchers, that it is a coherent notion and cannot be ruled out as impossible—especially when defined in terms of performance.

Another advocate of AI nativism is Tony Zador (2019). Zador argues that "in contrast to artificial neural networks, animals rely heavily on a combination of both learned and innate mechanisms." Today's neural networks "exploit only a tiny fraction of possible network architectures"; the practical (nativist) upshot is that AI researchers should look to biology for "more powerful, cortically-inspired architectures," which might better leverage innate knowledge to learn more effectively.

Just as an advantage of Margolis and Laurence's framing for cognitive science is that it avoids a trivial debate, the same is true in AI. The dispute between AI nativism and AI empiricism should not be about whether there are some innate features of the initial system. Marcus writes that "virtually every AI system contains lots of innate machinery that isn't acknowledged as such." His examples of "innate machinery" include "assumptions about how many layers should be included, how many units should be in each layer, what the input and output units should stand for, what activation function individual units should follow," and so forth. This is true so far as it goes, but counting this as evidence for nativism renders any AI system nativist, since any AI system will necessarily have unlearned features.

Instead, we can focus on the debate on a more substantive question: whether these unlearned features specify different innate mechanisms, states, and processes for different domains.

II. Arguments for and against necessity nativism

Given this target state, and different questions of what is possible versus what is practical, we can define different positions. In AI, *necessity nativism* is the claim that necessarily, a human-level AI system will be a nativist system. On this view, an AI system that is capable of general intelligence (possibly after extensive learning from data) will require nativist machinery—many distinct types of mechanisms, states, and processes for acquiring 'intelligence' traits, with different systems of acquisition operating across different domains.

The denial of necessity nativism is *possibility empiricism*. This is the claim that it is possible for a human-level AI to be an empiricist system. On this view, an AI system that is capable of general intelligence (possibly after extensive learning from data) does not require nativist machinery. Instead, it can have few distinct types of mechanisms, states, and processes for acquiring "intelligence" traits, with the same systems of acquisition operating across different domains.

How should we understand the strength of the *necessity* or *possibility* in these claims? I doubt that any necessity nativist would want to claim that nativist systems are metaphysically

necessary.[12] I suspect that the best way to understand the necessity in necessity nativism is something like nomologically necessary, or perhaps nomologically necessary across a restricted range of reasonable circumstances.

Practical nativism is a weaker position than possibility nativism. It is the claim that, in some sense, a nativist system is *more practical* for achieving human-level AI. (Practical nativism follows from necessity nativism, but not vice versa). "More practical" can be cashed out in a number of ways, with a number of different practical considerations. For example, the practical nativist can claim that building nativist systems will lead to human-level AI more quickly; or that nativist systems will be less data-hungry, or more reliable, or safer. *Practical empiricism* is the analogous position.

While I've drawn a distinction between "necessity" and "practicality," at the deepest level there will be a spectrum from widest metaphysical possibility to extremely narrow practical engineering constraint feasibility. Especially if we bake in some reasonableness constraints into "necessity," it might shade into "practicality." We should be clear that there really is a spectrum, and that for any given argument, we should be explicit about the background conditions we are evaluating for.

What kind of arguments might we give for each of these positions? We can start with arguments for necessity nativism and possibility empiricism.

Necessity nativism: the argument from human nativism

One argument for necessity nativism appeals to the (alleged) truth of nativism about the human mind. Humans are, arguably, the only general intelligence that we know about—or at least most powerful. If the human mind is nativist, this may be evidence that nativism is necessary for general intelligence. However, the inference from human nativism to necessity nativism is not very strong. Human innate machinery might be needed for creatures with our data, capabilities, and needs. Most evidence for nativism points, in various ways, to the limited data that children have access to, and their needs to learn quickly from this data. However, AIs will have data we do not have, capabilities that we lack, and needs different from ours.

Consider the question of data. A key argument for human nativism is the poverty of the stimulus argument, which appeals to the amount and types of data that humans have available to them. Here is a characterization of Chomsky's argument:

[12]A counterexample to metaphysical necessity nativism would be an astronomically large "empiricist" Blockhead that "learns" everything it needs to know by encountering exactly the right dataset—the one that lists all of the "right" actions in every task it will ever face. (Perhaps this counterexample could be excluded by requiring that it learn robustly in a variety of environments.)

1. If children were empiricist learners, then the data available to them would be too impoverished—they would not reliably arrive at the correct grammar for their language.
2. Children do reliably arrive at the correct grammar for their language.
3. Children are not empiricist learners.

By "empiricist learner," I mean (following Margolis and Laurence) one with "any innate domain-specific knowledge or biases to guide her learning and, in particular . . . any innate *language-specific* knowledge or biases" (222). Chomsky argues that without these language-specific knowledge or biases, the relevant data is not available to children.

This argument would apply if you were trying to build a machine that would be exposed to the same kind and amount of primary linguistic data as a child is. This argument would suggest that such a system would, like a human child, require language-specific knowledge in order to learn. However, there is no reason to think that we would need to limit the data available to an artificial system in this way, and thus no reason to think that this argument would apply. As we have seen, this argument relies in part on the fact that there are certain kinds of data that children do not get: for example, children don't get instances of language that are labeled as not grammatical. But this kind of data could be supplied in AI. And of course, the total amount of data supplied to AI is often greater than what is supplied to a child.

Call this "big data empiricism": the view that just because something is *not learned* by humans without innate machinery, does not necessarily show that it is *not learnable at all* without innate machinery, because more data can make the key difference. In AI, the stimulus is not necessarily poor: while children might not have enough data in their primarily linguistic data to learn the right grammar, an AI that needed to learn language could be given more data, and different data.

In addition, there are other relevant differences between AIs and humans. For one thing, humans have to learn quickly in order to survive in a hostile world. In some cases, we may have innate machinery not because this machinery is strictly necessary, but because it makes learning quick, reliable, and efficient.[13] But an artificial system need not learn things as quickly and reliably as we do.

These considerations apply beyond language acquisition. While the poverty of the stimulus is perhaps most widely known from language acquisition, poverty of the stimulus arguments are ubiquitous in nativist theorizing. For example, one might give a poverty of the stimulus argument

[13] Margolis and Laurence (2013); 11

about moral principles, or about the principles of intuitive biology or intuitive physics. But as with language, the available data may be far richer for artificial systems than it is for children.[14]

Big data empiricism can sometimes be used not just to eliminate nativist elements in human systems, but also to eliminate nativist elements in existing AI systems. For example, AlphaZero can use large amounts of computational power to learn across several different domains using the same architecture, something that was impossible for past game-playing systems. Gary Marcus complains that DeepMind's Go-playing system (AlphaGo) is not as empiricist as DeepMind claims, since it uses built-in convolutional layers to exploit the domain-specific fact that many patterns in Go are translation invariant.[15] It is true that many deep learning systems use convolutional layers, and that these layers in effect encode a nativist approach to spatial structure. However, shortly after AlphaGo, DeepMind showed that the same network architecture be used to play chess and shogi as well, which are *not* translationally invariant.[16] Thus, convolution ended up not being "domain-specific" to Go, and it was not necessary to tweak the architecture to fit each game. No domain-specific features were necessary. Or consider the ability of GPT-3 to generate grammatical sentences. It was not endowed with any unlearned principles of grammar, but with enough data and computational power, it generates mostly well-formed sentences.[17] The success of systems like AlphaZero and GPT-3 illustrates the power of big data empiricism: because of the opulence of the stimulus in AI, establishing necessity nativism requires more than poverty of the stimulus arguments.

Instead, necessity nativism requires impossibility arguments. Impossibility arguments allege that something is not learnable from *any* amount of data without some nativist machinery. As Marcus puts it, "With the right initial algorithms and knowledge, complex problems are learnable (or learnable given some sort of real-world constraints on computation and data). Without the right initial algorithms, representations and knowledge, many problems remain out of reach" (9).

Necessity nativism: the many-tasks argument and learnability

[14]For examples of general poverty of the stimulus arguments, see Mikhail 2008, "The Poverty of the Moral Stimulus," and Zaitchek and Samet (2017).

[15] "Many aspects of their system follow both from previous studies of computer Go (and game playing in general) and from the nature of the problem itself . . . artfully placed convolutional layers allow the system to recognize that many patterns on the board are translation invariant."

[16]Silver et al. (2017): "Go is well suited to the neural network architecture used in AlphaGo because the rules of the game are translationally invariant (matching the weight sharing structure of convolutional networks). . . Chess and shogi are, arguably, less innately suited to AlphaGo's neural network architectures."

[17]For a survey of the syntactic capabilities of deep learning models, see Linzen, Tal, and Marco Baroni. 2021. "Syntactic Structure from Deep Learning." Annual Review of Linguistics 7 (1): 1760100425. https://doi.org/10.1146/annurev-linguistics-032020-051035.

Impossibility arguments are often known as *learnability* arguments. A learnability argument for necessity nativism takes the following general form:

1. Capacity C is required for human-level AI.
2. Capacity C cannot be learned from data using domain-general empiricist learning mechanisms, and instead requires innate machinery N.
3. Therefore, human-level AI requires innate machinery N.

The many-tasks argument is a version of this master argument. Marcus is giving a many-tasks argument in the passage we saw earlier: "many different types of tasks will have their own innate requirements." Here is a formalization of this argument.

1. Human-level intelligence involves performing tasks T.
2. Requirements: Tasks T require capacities C.
3. (from 2, 3) Human-level intelligence requires capacities C.
4. Learnability: Capacities C cannot all be learned from data using domain-general empiricist learning.
5. If these capacities cannot be learned in this way, they require domain-specific innate machinery N.
6. Conclusion: Human-level intelligence requires domain-specific innate machinery N.

How does this argument fare? Each of the premises faces significant challenges. Premise 2, the requirement premise, is often very difficult to establish. Maybe it's not the case, as Marcus says, that syntactic tree manipulation operations will be required for language understanding. Of course, if success were defined as "doing the task in the same way that humans do it," then it's almost definitional that human-like representations will be necessary. But if task success is defined only in terms of human-level (or greater) performance, it's not guaranteed that success requires human-like representations. In fact, the history of AI shows that we are often wrong about what capacities are required for certain tasks—computers, with their unique advantages, often find extremely surprising and inhuman ways to perform tasks. In general, we should be skeptical of our ability to know what is required for a given task.[18]

The learnability premise is crucial. In order to establish it, one must give a *learnability argument*. Examples of learnability arguments in cognitive science include Fodor's argument that primitive concepts simply cannot be learned, and Goodman's argument that hypotheses formulated in terms of "grue" (as opposed to "blue") can never be ruled out from observations. In these cases,

[18]Hofstadter (1979) speculates: "There may be programs that can beat anyone at chess, but they will not be exclusively chess programs. They will be programs of general intelligence." Douglas Hofstadter (quoted in Weber 1996): "My God, I used to think chess required thought… Now, I realize it doesn't."

the argument is not that children don't have *enough* data. Rather, it's that without the right innate starting point, *no* amount of data would be enough.

Are there certain "intelligence" traits that simply cannot be learned from data, and which would make a nativist initial state necessary? Necessity nativism raises the question: how much could we have established about our own innate structure from *learnability* arguments, rather than from poverty of the stimulus arguments? How much of our own innate machinery is in principle *necessary*, as opposed to contingent to the way that we evolved—as creatures who must learn quickly from limited data?

Some cognitive scientists have used learnability arguments to argue for nativism about concepts. I will set aside Fodor's concept nativism, and focus on arguments by nativists like Carey and Spelke that certain "core" concepts cannot be learned. For example, Carey thinks this is true of concepts like "object" and "agent":

> [L]earnability considerations also argue that the representations in core cognition are the output of innate input analyzers. If the capacity to represent individuated objects, numbers, and agents are learned, built out of perceptual and spatiotemporal primitives, then there must be some learning mechanism capable of creating representations with conceptual content that transcend the perceptual vocabulary. In the second half of [*The Origin of Concepts*], I offer Quinian bootstrapping as a mechanism that could, in principle, do the trick, but this type of learning process requires explicit external symbols (words, mathematical symbols), and these are not available to young babies. Associative learning mechanisms could certainly come to represent regularities in the input, such as that if a bounded stimulus disappeared through deletion of the forward boundary behind another bounded stimulus there is a high probability that a bounded stimulus resembling the one that disappeared will appear by accretion of the rear boundary from the other side of the constantly visible bounded surface. But these generalizations would not be formulated in terms of the concept object. There is no proposal I know for a learning mechanism available to non-linguistic creatures that can create representations of objects, number, agency, or causality from perceptual primitives.[19]

One route to establishing necessity nativism in AI would be to formalize and extend such learnability arguments, such as the one suggested by Carey here. It is not currently clear that such arguments can be made to work in AI, but there is clearly much more to be said. For one thing, Carey is only maintaining that there are no learning mechanisms available to young babies, who

[19] Carey (2011), "Precis of Origin of Concepts."

lack explicit external symbols; perhaps AI techniques could leverage symbols to create domain-general learning mechanisms. In addition, a big data empiricist might wonder whether more data could make the difference (though Carey seems to think not). Such arguments would need substantial additional assumptions if they are to be used to support nativism about AI.

Possibility empiricism: evolutionary arguments

I now turn to arguments for possibility empiricism. A key motivation for possibility empiricism is that we know of at least one process that has resulted in general intelligence, starting "from scratch"—namely, biological evolution. Can this motivation be leveraged into an argument for possibility empiricism? The key issues will be the extent to which we can think of evolution as a form of learning (or as a process that can be recapitulated via learning), and how we think about evolution-inspired methods, especially architecture search, in AI.

Evolutionary processes lead to human-level intelligence; an evolutionary empiricist may hold that we can recapitulate these evolutionary processes as a form of domain-general AI technique. As Lake et al. (2016) explain this proposal (which they do not endorse): "The human brain effectively benefits from even more experience through evolution," and so deep learning may try "to capture the equivalent of humans' collective evolutionary experience."[20]

According to one kind of evolutionary empiricism, if certain intelligence pre-requisites can be evolved, they can be learned.

Evolution as learning

1. Evolution plus learning leads from a domain-general initial state to general intelligence.
2. Evolution can be recapitulated as learning.
3. If (1) and (2), then learning from a domain-general starting point can achieve general intelligence.

-

Learning from a domain-general starting point can achieve general intelligence.

But one might worry that evolution cannot be recapitulated as learning without trivializing the notion of learning. Gary Marcus objects that thinking of evolution as a form of learning weakens learning so much that it:

> . . . encompasses literally everything on either side of the debate, from what Locke had in mind to whatever Plato and Chomsky and Pinker had in mind, and everything in between. Relabeling the debate doesn't resolve the issues, either; one might as well just use the term

[20] While this form of evolutionary empiricism is couched in terms of deep learning, it could easily be generalized to other forms of learning-driven AI research.

> learning to refer to all change over time, regardless of mechanism, and count rock formations as the product of learning, too. Evolution (whether through natural selection or simulated artificial techniques) is a means towards building machinery with embedded prior knowledge, not an alternative to prior knowledge.

As a linguistic point, in machine learning, evolutionary processes tend to be counted as learning. But still, we should ask whether there is a deep sense in which evolution can be compared to learning, and thus whether there is reason to think it can be recapitulated as learning. One complication is that evolution operates over multiple systems.

To be sure, concepts of learning in psychology presuppose that learning is done by an individual, so evolution would obviously not count as learning. For example, De Houwer, Barnes-Holmes, & Moors (2013)[21] define learning as "changes in the behavior of an organism that are the result of regularities in the environment *of that organism."*[22] But we can count evolution as learning if we extend the scope of the process across several generations of organisms: changes that are the result of regularities in the environment of organisms.

To be sure, this kind of "learning" doesn't fit many cognitive science nativists' definition of learning straightforwardly. However, if evolution selects for the representational capacities of creatures, and those representational capacities must accurately track features of the world (how agents move, how objects behave, how scenes are usually lit) in order to promote fitness, then this process is a process of learning, in that it ensures a better fit between our representations and the world. The "evidence" that evolutionary processes respond to is very coarse-grained;[23] it only tracks reproductive fitness. But in doing so, it can indirectly track the fitness of representations to the world.[24]

It's helpful here to distinguish between a narrow sense of learning, which is learning *given an architecture*, and a broader sense, which can include architecture search. As an example, consider neural networks. In machine learning, the "learning" that neural networks do usually refers to supervised learning: adjusting connection weights between nodes in response to training

---

[21] Similarly, the authoritative textbook on learning in psychology, *Principles of Learning and Behavior* (Domjan 2010) says that "Learning is an inferred change in the organism's mental state which results from experience and which influences in a relatively permanent fashion the organism's potential for subsequent adaptive behavior."

[22] De Houwer, Jan, Dermot Barnes-Holmes, and Agnes Moors (2013). "What is learning? On the nature and merits of a functional definition of learning."

[23] As Sutton and Barto (2018) point out, it uses very coarse-grained evidence: it is indifferent to which states of the world any creature passes through. But it doesn't seem like there's a principled way to exclude this as an "evidential" process, even if it is very coarse-grained with respect to the batches of evidence it updates on.

[24] But perhaps it will *not* be true that we can see our innate machinery as "learned" if it is a mere byproduct, a spandrel, or simply a fluke. However, this is less likely to be true given that our innate machinery, or at least much of it, is shared with other animals.

data, in order to better approximate some function. To take a canonical and simple example, a single-layer neural network might be trained to represent the AND function—to only activate when both of its input nodes are activated. This is done by adjusting the weights between input and output layer. This is learning in a narrow sense: out of a set of possible functions, the best one for making sense of the input data is learned.

Minsky and Papert (1969) pointed out that a single perceptron is incapable of learning the XOR function—it is simply outside of the scope of functions that it can learn. This is, in its way, a very simple learnability argument for a given architecture. However, a multi-layer network *is* able to learn XOR. Now suppose we have a process that involves selecting the best architecture for the problem—which results in an evolution from a single-layer network to a multi-layer network—and then training the best architecture on the problem at hand. This process contains learning in both the broad and narrow sense.

This is a toy example, but it matches much of AI practice. In AI, before "narrow" supervised learning takes place, many key decisions must be made about the architecture to be used. Often these choices are made by hand, with one or more people trying out many different architectures (often these people are underlings, hence the sardonic term "grad student descent"). But these trial and error process can be automated. For example, approaches like the neuroevolution paradigm use artificial evolution to evolve different networks, which are then trained on data.[25]

System individualization matters here. It's true that the selected system, the multi-layer network, can only succeed because it has the right set-up before "narrow" learning. But the overall process does not start with nativist machinery. The process as a whole can be viewed as an empiricist approach to succeeding on XOR.

This suggests a master argument for evolutionary empiricism: For any given innate starting points that are necessary, these innate starting points can be "learned" through architecture search rather than built in by hand. We might call this view *architecture-search empiricism.*

In response, a nativist might suggest that architecture search really supports nativism, by requiring that nativist architectures must evolve. *Architecture-search nativism* says that a process of artificial evolution will, on its waypoint to producing intelligent creatures, necessarily create nativist entities — perhaps entities whose architectures are relevantly similar to ours. Why would anyone hold that nativist waypoints are necessary? One might believe this thesis if one believes that, as Spelke and Blass (2017) put it, "Core knowledge captures fundamental properties of space,

[25]Such, Felipe et al.. Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning. arXiv:171206567 [cs] [Internet]. 2017 Dec 18; Available from: http://arxiv.org/abs/1712.06567

objects, and agency." One would have to further believe that they are so fundamental that you can't get to intelligence without a system that first has Spelke-style core concepts.

Substantively, architecture-search nativism and architecture-search empiricism are not in conflict with each other. The key difference is the verbal issue of whether architecture search counts as a form of learning (empiricism) or as something else (nativism). Such processes are almost certainly in *principle* available, since evolution is a proof of concept that a causal process can produce the right innate machinery for general intelligence through architecture search. If architecture search counts as a domain-general learning process, then it seems that possibility empiricism about human-level AI is correct. If it does not, then the case is significantly weaker.

I will not try to adjudicate the verbal issue of whether architecture search counts as learning, nor are we in a position to settle the substantive issue of whether there is a route to human-level AI that does not have nativist architectures as a waypoint. If the answer to either question is yes, possibility empiricism is almost certainly correct. If the answer to both questions is no, necessity nativism is almost certainly correct. As a result, the dialectical situation perhaps slightly favors possibility empiricism. But in any case, it is clear that these two issues concerning architecture-search nativism and empiricism will be among the key points in adjudicating the question between nativism and empiricism in AI.

III. Arguments for practical nativism and practical empiricism

What about practical nativism? Practical nativism is a weaker position than necessity nativism. Practical nativism is compatible with possibility empiricism: even if it is possible in principle to learn the necessary ingredients of intelligence from an empiricist starting point, it may be in practice more desirable to build machines with nativist starting points. Arguments for practical nativism or practical empiricism appeal to the various advantages of a given approach: these advantages can include engineering feasibility and the properties of the final system.

I think that the crucial considerations that will drive which approach is more practical will be: how feasible it is to directly encode the "right" innate machinery, and how efficient evolutionary techniques are. The easier it is to directly specify the right innate machinery, the more practical it is to do this instead of trying to evolve it, which would favor practical nativism. On the other hand, the more efficient evolutionary techniques are, the better they are as an alternative to this project, which would favor practical empiricism.

Practical nativism: arguments from efficiency and imitation

For example, even if an evolutionary argument establishes possibility empiricism, one might argue that it is wasteful to have machines recapitulate the learning of evolution. For example,

Marcus writes that "an unthinking commitment to relearning everything from scratch may be downright foolish, effectively putting each individual AI system in the position of having to recapitulate a large portion of a billions years of evolution." Baldassare et al. (2017) argue that evolutionary methods might still be impractical, due to the immensity of the search space:

> …one thing that the enormous genetic algorithm of evolution has done in millions of years of the stochastic hill-climbing search is to develop suitable brain architectures. One possible way to attack the architecture challenge, also mentioned by Lake et al., would be to use evolutionary techniques mimicking evolution. We think that today this strategy is out of reach, given the "ocean-like" size of the search space.[26]

The practical empiricist has a reply here: it's not necessary to recapitulate all of evolution. Perhaps evolutionary approaches set up by humans can be considerably more efficient than evolution. Evolution does not directly select for intelligence, whereas evolutionary methods can select for a variety of things related to intelligence.[27]

However, the practical nativist can reply that in any case we might as well make use of what we know: that the general intelligences we do know about do come with equipped with innate machinery. Given that engineering intelligence is still exceedingly difficult, it would be foolish to not try look for inspiration from natural intelligences. Consider the following hope from Spelke and Blass (2017): "Because human infants are the best learners on the planet and instantiate human cognition in its simplest natural state, a computational model of infants' thinking and learning could guide the construction of machines that are more intelligent than any existing ones." The slogan here is: why not imitate the one intelligent system that we *do* already know about? This system is likely nativist; why not try nativist approaches in AI?

Practical nativism: arguments from data efficiency

[26]Baldassarre, Gianluca et al. (2017). "The architecture challenge: Future artificial-intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction." The Behavioral and brain sciences. 2017;40:e254.

[27] As with so many foundational issues in AI, this point was anticipated by Turing (1950) in "Computing Machinery and Intelligence": "We have thus divided our problem into two parts: the child-programme and the education process. These two remain very closely connected. We cannot expect to find a good child-machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, by the identifications: Structure of the child machine = Hereditary material; Changes = Mutations; Natural selection = Judgment of the experimenter. One may hope, however, that this process will be more expeditious than evolution. The survival of the fittest is a slow method for measuring advantages. The experimenter, by the exercise of intelligence, should be able to speed it up. Equally important is the fact that he is not restricted to random mutations. If he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it."

In addition, the practical nativist can argue that innate machinery will allow for learning with less data. Humans are, in some domains at least, much less data-hungry than current state-of-the-art machine learning methods. Google Translate and GPT-3 have seen far more text than any human has; AlphaZero has played millions more games than Lee Sedol. Our data efficiency may be due in large part to our innate machinery, which "homes in on" certain hypotheses. If data efficiency is a key desideratum for practicality, then this might favor practical nativism.

That said, there are tricky methodological questions about comparing the "data" that AI gets and the "data" that human infants get. As DeepMind's Adam Santoro (2019) puts it,

> Comparisons are often made. . . between animals and ANNs [artificial neural networks] learning on supervised datasets. As the comparisons go, these ANNs need millions of labelled examples; vast quantities more than animals receive by the time they exhibit impressive behaviours. Unfortunately this is an apples-to-oranges comparison. Animals receive a glut of extremely high-quality data that reveals orthogonal factors of variation, unlike the static sets of images, which are filled with spurious correlations that entrap ANNs. No amount of training steps can make up for impoverished data. . . We have zero proof that (potentially embodied) ANNs learning on an equivalently rich stream of data cannot exhibit behaviours similar to animals. We only have proof that ANNs learning from an abundance of massively impoverished data do not.

For this reason, it is not clear that data efficiency is a strong argument in favor of practical nativism.

Practical empiricism: arguments from ignorance and laziness

We've learned a lot about the innate initial state of the human infant. For example, we know that by two months, children expect that objects cohere, are solid, move continuously, and so forth. We know that children have some innate knowledge of possible grammars. But do we know how exactly to encode this knowledge computationally? The way these expectations are embedded might actually be extremely complex, and not something that we can encode directly (at this stage of our knowledge).

Call this the argument from ignorance: we just don't currently know how to skip straight to the initial state that evolution got us. And in the meantime, trying to do so may be counterproductive, as we add constraints that are not helpful. This is the thought behind the oft-quoted—and probably apocryphal—quip by speech recognition pioneer Fred Jelenik: "Every time I fire a linguist, the performance of our speech recognition system goes up." Richard Sutton has written that the "bitter lesson" of 70 years of AI research is "general methods that leverage computation are ultimately the most effective, and by a large margin":

> Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. . . . the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation.

This ignorance could change. But the practical empiricist responds to the practical nativist simply by saying: okay, show me *exactly* what you mean by programming more "innate machinery" into machines. Until we know, we can supplement the argument from ignorance with an argument from laziness: perhaps training is easier than programming the right things in. And if empiricism is even possible, it may be easier to come up with the right learning algorithm and the right data to get there.

Practical empiricism: arguments from ambition

Evolution doesn't find the optimal solution: it finds the easiest solution available. In encoding the innate machinery that evolution gave us, the practical empiricist argues, we might be insufficiently ambitious. Why not have artificial systems learn, or evolve, better solutions? Such solutions might not make reference to our own innate concepts of "object" or "agent," but why think these are the best concepts for carving up the world, for every system? (Note that the argument from interpretability, below would reply that a system with different concepts than these might be incomprehensible to us and harder to align with our values.)

Against this, the nativist might argue that core knowledge is not merely the best hack evolution got together, but rather reflects fundamental constraint on cognition, because it captures deep regularities about the world. For example, Spelke and Blass write that ". . . core knowledge captures fundamental properties of space, objects, and agency." On the other hand, if this is the case, the empiricist can reply that we can let systems learn these fundamental properties themselves. Whether this kind of learning is possible, as we have seen, depends on the strength of learnability arguments.

Practical nativism: arguments from explainability

Most of these concerns have been about how long or difficult or costly the path to HLAI might be. But another key consideration is the *kind* of HLAI that is created, and the risks or benefits associated with this system.

A slogan for practical empiricism might be: "empiricist approaches are more powerful because they enable solutions that we could humans never dream of" (see: the argument from

ambition). But this highlights the very feature of empiricist AI that would make it potentially more dangerous, and suggests a transparency argument for nativism. Empiricist approaches allow for the possibility of systems that solve the same problems that we do, but without utilizing our (partially innate) concepts of agency, objecthood, and causation. If our ability to understand other agents depends on their sharing these concepts, then it might be impossible for us to understand the actions of an HLAI that does not share these concepts. In contrast, a route to human-level AI that begins with human-like innate machinery, the nativist argues, will be more likely to produce human-level AIs that are interpretable to us. And interpretability conducive to many pragmatic and ethical goals, such as preventing bias and unfairness and ensuring safe and value-aligned behavior. In other work[28], I trace the ways in which explainability is a precondition for fairness and for safety. But for now we simply note that explainability is *prima facie* a very desirable trait for AI systems to have; it is difficult to ensure the fair and safe behavior of a system that is inscrutable to you.

What is it for an AI system to be explainable? Lipton (2016) points out that the field of explainable/interpretable/transparent AI uses these terms in many different ways, associated with different reasons we might care about whether AI is explainable/interpretable/transparent. For our purposes, I use the term "explainability" with the following meaning: An AI system is explainable to the extent that we are able to give humanly comprehensible reasons for (or at least explanations of) its predictions and actions.[29]

To get a grip on this, consider how current AI systems *fail* to be explainable. Due to the black-box nature of deep neural networks, are often surprised by the behavior of these systems, since we don't really understand the function that they have learned. The problem of "adversarial examples" makes this vivid.[30] An adversarial image is one which is misclassified by a machine learning model, even though it is similar (or even indistinguishable) to the human eye from one which the model classifies correctly. Such examples reveal that the function that that the model uses to classify images is very different from whatever the human "classification function" is. Understanding why the model manifests this strange behavior this way is very difficult, even though we in some sense have a full knowledge of the model. But simply possessing a long list of connection weights does not enable us to understand or predict this surprising behavior.

Although deep neural networks are often extremely complex, the reason for this lack of understanding is not the complexity of AI systems *per se*; although often associated with inexplicability,

---

[28] [redacted for review] (ms) "Transparency, fairness, and safety in artificial intelligence"

[29]In stipulating that the reasons / explanations be *humanly comprehensible*, I'm following Doshi-Velez and Kim (2017), who gloss explainability as "the ability to explain or to present in understandable terms *to a human."* This means that human cognitive limitations are an important part of what makes a system interpretable. This definition also applies whether or not one thinks that artificial systems can have "reasons" for action, as opposed to explanations of their actions.

[30]See Goodfellow et al. (2014) for the canonical paper on adversarial examples.

complexity is neither necessary nor sufficient for inexplicability. For one thing, simple systems can be unexplainable. For instance, a simple decision tree that employs "unnatural", alien engineered features is difficult to explain. At the same time, complex systems can be explainable: for example, we are able to understand some features of human behavior and of economics using high-level explanations, even though human beings and economies are incredibly complex systems.

The problem is, instead, alien-ness. Our innate endowment has given us a certain way of carving up the world: into objects, agents, causes, intentions, and so forth. These "core concepts" are useful for creatures with our particular limitations and needs. They also allow us to understand the behavior of other creatures who share these concepts. But it is very unlikely that every possible intelligence will need to share these same concepts in order to successfully navigate the world. In fact, it is likely that artificial systems with more memory, different developmental trajectories, and more data, need not hit upon these same abstractions in the course of learning. To the extent that they do not, it will be difficult for us to understand these machines—and thus difficult to ensure that their behavior is safe or fair.

Thus, a wide solution space is part of the power, but also the inexplicability, of empiricist methods. It allows empiricist methods to employ concepts that are foreign to us. Building in human domain-specific starting points is a crucial way of constraining the solution space and finding explainable solutions. As Ilyas et al. (2019) put it, "attaining models that are robust and interpretable will require explicitly encoding human priors into the training process."[31] This is, in the language of this paper, a statement of practical nativism motivated by explainability.

Of course, nativist AIs can *still* be unsafe, if we are unable to install the "right" human priors, or for other reasons. But the explainability nativist argues that there is at least a greater chance of explainable AI through nativist means, because empiricist approaches are likely to hit upon alien and inhuman solutions.

IV. Conclusion

In this paper, I have developed nativism and empiricism as a framework for thinking about artificial intelligence. Linking ongoing debates in AI research to the historical conversation between nativists and empiricists helps us to frame them as debates about the nature of intelligence in general. How *contingent* is it that we have the native machinery that we do? Is this the result of some general of "law of intelligence," or just some hacks that evolution hit upon for creatures like us, with our data?

[31] Ilyas, Andrew, et al. 2019. "Adversarial Examples Are Not Bugs, They Are Features." ArXiv:1905.02175 [Cs, Stat], August. http://arxiv.org/abs/1905.02175.

We've seen that, on the issue of necessity nativism versus possibility empiricism, evolutionary empiricism gives a strong case for possibility empiricism. This outcome may be complicated by which side gets to claim victory if architecture search is necessary for human-level AI. Beyond this verbal issue, there remain substantive questions about whether innate machinery is a necessary "waystation," and about whether there are learning-based alternatives to architecture search. It may be that just as this debate forces us to sharpen our concept of learning, it will require us to sharpen our concept of architecture search.

On the practical side, there is a tradeoff between the advantages of encoding innate machinery directly, and the advantages of evolving or learning it. Right now, there is no clear answer to the question of practical nativism versus practical empiricism in AI, though I've argued that the desideratum of explainability gives us strong normative reasons to favor practical nativism. For now, we can expect that different researchers will continue to take different bets on nativist and on empiricist research programs, and a thousand flowers will bloom.

Works cited

Baldassarre, Gianluca, Vieri Giuliano Santucci, Emilio Cartoni, and Daniele Caligiore. 2017. "The Architecture Challenge: Future Artificial-Intelligence Systems Will Require Sophisticated Architectures, and Knowledge of the Brain Might Guide Their Construction." *The Behavioral and Brain Sciences* 40: e254. https://doi.org/10.1017/S0140525X17000036.

Barsalou, Lawrence W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22 (4): 577–660. https://doi.org/10.1017/S0140525X99002149.

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.

Buckner, Cameron. 2018. "Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks." *Synthese* 195 (12): 5339–72. https://doi.org/10.1007/s11229-018-01949-1.

Cohen, Jonathan. 2015. "Perceptual Representation, Veridicality, and the Interface Theory of Perception." *Psychonomic Bulletin & Review* 22 (6): 1512–18. https://doi.org/10.3758/s13423-014-0782-3.

De Houwer, Jan, Dermot Barnes-Holmes, and Agnes Moors. 2013. "What Is Learning? On the Nature and Merits of a Functional Definition of Learning." *Psychonomic Bulletin & Review* 20 (4): 631–42. https://doi.org/10.3758/s13423-013-0386-3.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." ArXiv Preprint ArXiv:1702.08608.

Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1997. *Rethinking Innateness: A Connectionist Perspective on Development*. Reprint edition. Cambridge, Mass.: A Bradford Book / The MIT Press.

Hoffman, Donald D., Manish Singh, and Chetan Prakash. 2015. "The Interface Theory of Perception." *Psychonomic Bulletin & Review* 22 (6): 1480–1506. https://doi.org/10.3758/s13423-015-0890-8.

Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." ArXiv:1905.02175 [Cs, Stat], August. http://arxiv.org/abs/1905.02175.

Lake, Brenden M., and Gregory L. Murphy. 2020. “Word Meaning in Minds and Machines.” ArXiv:2008.01766 [Cs], August. http://arxiv.org/abs/2008.01766.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2016. “Building Machines That Learn and Think Like People.” *ArXiv:1604.00289 [Cs, Stat]*, April. http://arxiv.org/abs/1604.00289.

LeCun, Yann and Gary Marcus. “Does artificial intelligence need more innate machinery?” Debate hosted by NYU Center for Mind, Brain, and Consciousness. October 5, 2017. https://wp.nyu.edu/consciousness/innate-ai/

Linzen, Tal, and Marco Baroni. 2021. “Syntactic Structure from Deep Learning.” Annual Review of Linguistics 7 (1): 1760100425. https://doi.org/10.1146/annurev-linguistics-032020-051035.

Lipton, Zachary C. 2018. “The Mythos of Model Interpretability.” Queue 16 (3): 31–57.

Marcus, Gary. 2018a. “Deep Learning: A Critical Appraisal.” *ArXiv:1801.00631 [Cs, Stat]*, January. http://arxiv.org/abs/1801.00631.

———. 2018b. “Innateness, AlphaZero, and Artificial Intelligence.” *ArXiv:1801.05667 [Cs]*, January. http://arxiv.org/abs/1801.05667.

Marcus, Gary F. 2018. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT press.

Margolis, Eric, and Stephen Laurence. 2013. “In Defense of Nativism.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 165 (2): 693–718.

Minsky, Marvin, and Seymour Papert. 2017. *Perceptrons, Reissue Of The 1988 Expanded Edition With A New Foreword By Léon Bottou*. The MIT Press. https://mitpress.mit.edu/books/perceptrons-reissue-1988-expanded-edition-new-foreword-leon-bottou.

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

Ramsey, William, and Stephen Stich. 1990. “Connectionism and Three Levels of Nativism.” *Synthese* 82 (2): 177–205.

Samet, Jerry. 1987. “Troubles with Fodor’s Nativism.” *Midwest Studies in Philosophy* 10 (1): 575–594.

Samet, Jerry, and Deborah Zaitchik. 2017. “Innateness and Contemporary Theories of Cognition.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/innateness-cognition/.

Santoro, Adam. 2019. “Thoughts on ‘A Critique of Pure Learning’, Zador (2019).” Medium. October 17, 2019. https://medium.com/@adamsantoro/thoughts-on-a-critique-of-pure-learning-zador-2019-820a7dbbc783.

Shulman, Carl, and Nick Bostrom. 2012. “How Hard Is Artificial Intelligence? Evolutionary Arguments and Selection Effects.” *Journal of Consciousness Studies* 19 (7–8): 7–8.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. “A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play.” *Science* 362 (6419): 1140–44. https://doi.org/10.1126/science.aar6404.

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.

Spelke, Elizabeth. 1994. “Initial Knowledge: Six Suggestions.” *Cognition* 50 (1): 431–45. https://doi.org/10.1016/0010-0277(94)90039-6.

Spelke, Elizabeth S., and Joseph A. Blass. 2017. “Intelligent Machines and Human Minds.” *Behavioral and Brain Sciences* 40. https://doi.org/10.1017/S0140525X17000267.

Spelke, Elizabeth S., and Katherine D. Kinzler. 2009. “Innateness, Learning, and Rationality.” *Child Development Perspectives* 3 (2): 96–98. https://doi.org/10.1111/j.1750-8606.2009.00085.x.

Such, Felipe Petroski, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. 2017. “Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning.” *ArXiv:1712.06567 [Cs]*, December. http://arxiv.org/abs/1712.06567.

Sutton, Richard S. 2019. “The Bitter Lesson.” March 13, 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.

“The Mystery of Go, the Ancient Game That Computers Still Can’t Win | WIRED.” 2014. Accessed February 20, 2019. https://www.wired.com/2014/05/the-world-of-computer-go/.

Turing, Alan 1950. “Computing Machinery and Intelligence.” *Mind* LIX (236): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Weber, Bruce. 1996. “Mean Chess-Playing Computer Tears at Meaning of Thought.” *New York Times* 19.

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