Robert Long
Draft 3 February 2020
Work in progress. Comments, sharing welcome! Just let me know (rl2898@nyu.edu).

Nativism and empiricism in artificial intelligence

Introduction

In 2014, many researchers predicted that it would be at least a decade until machines could play Go as well as humans.[1] But just two years later the AI company DeepMind unveiled AlphaGo, a Go-playing system that surpassed all previous artificial and human players.[2] A successor, AlphaZero, plays superhuman Go, chess, and shogi.[3] AlphaZero achieved its superhuman performance by playing millions and millions of games against itself. It used reinforcement learning to search through the space of possible policies and find the winning ones. Go is over 2,000 years old, but after just four hours of training, AlphaZero played better Go than any system, human or machine, that has ever existed.

AlphaZero excited so many people because it learned all of its Go, chess, and shogi strategy, in some sense, from scratch. It lacked the kinds of game-specific, hand-crafted input 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. For this reason, DeepMind's engineers claimed that AlphaZero shows that "a general-purpose reinforcement learning algorithm can achieve, tabula rasa, superhuman performance across many challenging domains."[4] This is a deliberate allusion to John Locke, who argued that the human mind is a blank slate, or *tabula rasa*, with very little inbuilt knowledge about how the world works, and that we build all of our knowledge from experience.

The DeepMind research strategy is self-consciously *empiricist*: build a system with very little domain-specific knowledge, and let the system learn what it needs from data, using domain-general learning algorithms. The success of AlphaZero, and of learning techniques like deep learning and reinforcement learning more generally, has given AI a relatively empiricist bent in the last decade—at least in the way that AI researchers conceive of what they are doing.

Historically, the dispute between empiricists and nativists has been about the nature of human and animal intelligence. But as we can see, empiricist and nativist concerns now arise in artificial intelligence. This paper uses nativism and empiricism to pose questions about the nature of artificial intelligence: will the development of truly general artificial intelligence (AGI) require the construction of nativist systems, or can AGI 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 AGI is possible. That is, it is

[1] Levinovitz A. The mystery of Go, the ancient game that computers still can't win. Wired Magazine. 2014. Available from: https://www.wired.com/2014/05/the-world-of-computer-go/
[2] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017;550(7676):354.
[3] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science. 2018 Dec 7;362(6419):1140–4.
[4] Silver et al. (2018)

unlikely that nativist starting points are necessary for AGI, though there may be some role for nativist architectures at an intermediate stage. It remains an open question, however, whether empiricist or nativist approaches will be more successful in practice as a path to AGI.

I. Broadening nativism and empiricism for artificial intelligence

In cognitive science, nativists and empiricists seek to explain a central mystery: how do human beings, who are born so ignorant and helpless, come to be so smart? That is, they seek to explain the development of those psychological traits that are necessary for, or constitutive of, mature human intelligence. These traits reliably develop in 'normal' development, and one question is what the infant mind must be like in order for this development to be possible.

Consider linguistic competence. Noam Chomsky's theory of language acquisition is the *locus classicus* of modern nativism. Chomsky asked: given the data that children are exposed to—their '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 others' 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 a famous empiricist framework, 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]

Each of these theories are instances of more general approaches. Margolis and Laurence (2012) claim that nativists and empiricists fundamentally disagree about whether we should posit *many* different innate traits operating in *different domains* (nativism), or whether we should posit *few* innate traits that operate in a *domain-general* way (empiricism). According to Margolis and Laurence:

[5] Chomsky N. Aspects of the theory of syntax. Cambridge, MA: MITPress. 1965.
[6] Carey S. The Origin of Concepts. Oxford University Press; 2009. 609 p.
[7] Spelke E. 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).

> [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.
>
> [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.

These definitions rest a lot on the notion of a "domain." One might wonder how we are supposed to individuate domains. In cognitive science, people seem to use this notion while leaving it unexplicated, taking it as a given that there is some meaningful sense in which there is a domain of language acquisition, and that Universal Grammar encodes principles specific to this domain. But notice that a domain can be quite broad, relevant to almost all circumstances and environments. For example, Susan Carey speaks of the "object system," the "agency system," and the "number system." While these are (according to Carey) separate domains, with relatively encapsulated systems of acquisition and their own principles governing, they deal with very basic concepts that apply widely in the life of the organism—objects and agents are ubiquitous, after all.

Relatedly, one might wonder about the relationship between nativism and modularity—the claim that the mature mind is individuated into different encapsulated processes. Modularity and nativism, while related and often held in conjunction, can come apart. Nativism is a claim about the unlearned initial features of the mind, while modularity is a claim about the character of the mature mind. So it is possible to hold that the mature mind is modular, but that this modularity is developed after learning from an empiricist starting point. On such a view, the mind begins with very general learning systems which over time result in relatively encapsulated and domain-specific processes. On this empiricist-modularity view, differentiation is the result of the empiricist processes operating in a variety of domains.

Notice also that because this definition makes references to several different things—"mechanisms, states, and processes"—it allows for intermediate views where these come apart. For example, there could be a system in which there is only one domain-general learning process, but many domain-specific stores of prior information upon which this learning process operates. Consider a system in which all learning is done by the same Bayesian updating algorithm, but which starts out with very detailed prior information which encodes domain-specific information; certain hypotheses are favored in folk biology, others in vocabulary-learning, etc. For example, perhaps the hypothesis that a given word refers to round things gets assigned greater prior probability than the hypothesis that it refers to blue things. This system is empiricist in that it posits a single process, but nativist in that it posits detailed domain-specific initial states. If an "intermediate" theory like this ends up being right, we needn't lose any sleep over whether it is "more" empiricist or "more" nativist. Margolis and Laurence's characterization is still useful in drawing our attention to the issue of domain generality versus domain specificity on which nativists and empiricists often disagree.

Finally, one especially helpful feature of this framing is that it doesn't make the issue come down to whether anything at all is innate. Whether in cognitive science or in AI, everyone should agree that *something* must be innate. As Jerry Samet puts it, "Everyone agrees that learning requires that something be innate—even tabulas have some innate structure."[9]

[9]Samet J. Troubles with Fodor's nativism. Midwest Studies in Philosophy. 1987;10(1):575–594. p575

The key issue is instead whether humans (or perhaps non-human animals) are empiricist or nativist *systems*. A nativist system is one that is equipped with many distinct types of mechanisms, states, and processes for acquiring 'intelligence' traits, with different systems of acquisition operating across different domains. An empiricist system is one that is equipped with few distinct types of mechanisms, states, and processes for acquiring 'intelligence' traits, with the same systems of acquisition operating across different domains.

Formulating the issue in terms of empiricist and nativist systems allows us to unify the discussion of empiricism and nativism in cognitive science and AI. The issue in cognitive science is about whether *actually existing systems* (humans and non-human animals) are nativist or empiricist systems. Nativist or empiricist theories in cognitive science appeal to inference to the best explanation: given the data that children are exposed to, and the capacities that they develop, what is the character of the *actual* unlearned state, such that it explains this development?

In AI, nativism and empiricism are more naturally focused on what systems it is *possible* and perhaps *practical* to build. The issues concern what kind of initial state of an artificial system is necessary, or practical, for that system to be able to reach some final state, given some data. For example, you might wonder what you would need, or want, to build in to 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 is making a claim about what it is possible for general-purpose reinforcement learning algorithms to achieve.

Of course, artificial intelligence concerns itself with all kinds of target systems, including narrow 'task' AIs that only do one thing (play StarCraft, translate written text). In this paper I will primarily be interested in systems of *artificial general intelligence*. Of course, such systems do not yet exist. By artificial *general* intelligence, I mean 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. Or, as some people put it, an AGI would be able to flexibly pursue its goals in a wide range of environments. Crucially, I am indexing AGI to *performance* – on this usage, an AGI need not perform tasks in the same way that humans perform them. Analogously, AlphaZero counts as a superhuman Go player even though it doesn't play or learn Go in a human-like way (human beings don't learn Go by playing millions of games against themselves).

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, an AGI 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 an AGI 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.[10] 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 AGI. (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 AGI 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 it really is a spectrum and for any given argument we should be explicit about the background conditions we are evaluating for.

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 Gary Marcus is necessary for artificial general intelligence.

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 inbuild 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 AGI 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.

[10] 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.)

Another advocate of AI nativism is 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.

Avoiding trivial nativism

Recall that an advantage of the Margolis and Laurence framing for cognitive science was that it does not reduce nativism to an uninteresting "trivial" nativism. The same is true in AI: AI nativism should be a more substantive thesis than whether there are some innate features of the initial system. Although Marcus clearly believes in something more substantive than trivial nativism, at times his arguments only support trivial nativism. For instance, he 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 sort of "innate machinery" would render any AI system "trivially nativist," since any AI system has to specify these unlearned features.

By couching the debate in terms of domain-generality versus domain-specificity, my definition avoids trivializing the question in this way, instead focusing on the debate on a more substantive question: whether different tasks and different domains will require different innate assumptions. As it happens, Marcus also thinks the answer to this question is 'yes,' as he believes that "many different types of tasks will have their own innate requirements." Thus, he is an AI nativist in that he thinks that distinct capacities will arise from domain-general learning mechanisms and data alone. LeCun, by contrast, is empiricist in that he holds that these domain-specific requirements are either unnecessary, or can be learned via a domain-general process like deep learning or reinforcement learning.

II. Arguments for and against necessity nativism

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. The reason it's not very strong is illuminating. 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 Laurence and Margolis's (2001) construal of the poverty of the stimulus argument, as it is used in to establish nativism about language acquisition.

The Standard Poverty of the Stimulus Argument in language acquisition

1. An indefinite number of alternative sets of principles are consistent with the regularities found in the primarily linguistic data.
2. The correct set of principles needn't be (and typically isn't) in any pretheoretic sense simpler or more natural than the alternatives.
3. The data that would be needed for choosing among these sets of principles are in many cases not the sort of data that are available to an empiricist learner in the child's epistemic situation.
4. So if children were empiricist learners, they couldn't reliably arrive at the correct grammar for their language.
5. Children do reliably arrive at the correct grammar for their language.
6. Therefore, children aren't empiricist learners.

They continue, "by an *empiricist learner,* we mean one that…wouldn't have any innate domain-specific knowledge or biases to guide her learning and, in particular, wouldn't have any innate *language-specific* knowledge or biases" (222). Laurence and Margolis write that without these language specific knowledge or biases, "the decisive data just aren't available. The upshot is that when it comes to language, children aren't empiricist learners" (222).

Now, suppose that you were going to build a machine that would be exposed to the same kind and amount of primary linguistic data that the child was. 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 is often greater.

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.[11] 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

[11] Laurence and Margolis (2013), 'In defense of nativism'

about moral principles, or about the principles of intuitive biology or intuitive physics. But as with language, the stimulus may be far richer for artificial systems than it is for children.[12]

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.[13] 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.[14] 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. The success of AlphaZero 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

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 AGI
2. Capacity C cannot be learned from data using domain-general empiricist learning mechanisms, and instead requires innate machinery N
3. Therefore, AGI 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. General intelligence involves performing tasks T.
2. Requirements: Tasks T require capacities C.

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

[13] "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."

[14]"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." Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. 2017 Dec 5; Available from: https://arxiv.org/abs/1712.01815.

3. (from 2, 3) General 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: General intelligence will require domain-specific innate machinery N

How does this argument fare? Each of the premises faces significant challenges. Premise 2, the requirement premise, can be challenged. It is often very difficult to establish with certainty that a given task has a certain requirement—maybe it's not the case, as Marcus says, that syntactic tree manipulation operations will be required for language understanding. Of course, if success at a task is 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, then there's less reason to think 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.[15]

Secondly, we might dispute whether general intelligence involves performing certain tasks. However, I'll take it that often premise 1 is stipulative. Often, what people *mean* by AGI is defined in terms of task performance: by AGI they simply mean something that can play Go, translate languages, cooperate with other agents, track objects, and so forth.

Premise 4 is the key learnability premise. A *learnability argument* must be given for it. 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, the argument is not that children don't have *enough* data – 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 for the capacities that we have, rather than from poverty of the stimulus arguments? How much of our innate machinery is in principle necessary, as opposed to contingent to the way that we evolved—as creatures who need to learn quickly from limited data?

Some cognitive scientists have used learnability arguments to argue for nativism about concepts. I will set aside Fodor's notorious 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

[15]Douglas 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."

> 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.[16]

One route to establishing necessity nativism in AI is 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 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 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.

It's not surprising that the modern debate between nativists and empiricists takes place with evolution in the background. Evolution provides a plausible mechanism by which innate structures come into existence, thereby making nativism much more plausible; it also provides a wider perspective on learning. As Samet and Zaitchek (2017) note: "A more enterprising Empiricism might have noted that evolutionary theory commits us to the idea that whatever is innate in us was, at least in one sense, shaped by experience. Experience here would be ancestral experience, not the experience of the individual subject…the range of 'learning from experience,' the Empiricist's core commitment, would simply be extended to cover not only individual learning but species-based learning as well."

Here we see the potential for an evolutionary empiricism to undercut necessity nativism. 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,"

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

and so deep learning may try "to capture the equivalent of humans' collective evolutionary experience.[17]

According to this kind of evolutionary empiricism, if certain intelligence pre-requisites can be evolved, they can be learned. The simplest way of arguing for this relies on the idea that evolution is learning:

1. Evolution is a process of learning that results in the initial state of the human organism.
2. If evolution is a process of learning that results in the initial state of the human organism, then learning from an empiricist starting point can achieve general intelligence.
3. Learning from an empiricist starting point can achieve general intelligence.

Some AI nativists want to draw the lines of the debate so that evolutionary processes are excluded from vindicating the empiricist position. For example, Marcus complains that thinking of evolution as learning trivializes the notion of learning, so 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 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.

To be sure, defining learning only as "all change over time" is not helpful. But in machine learning, evolutionary processes tend to be counted as learning. Is there a principled way to count (some processes of) evolution as learning? I believe we can do better than counting *any* causal process as a process of learning, as Marcus suggests.

The distinction between evolution and learning is considerably thornier in AI nativism versus AI empiricism than it is in the "classic" debate. That's because in the "classic" debate, there is a relatively well-defined system about which nativists and empiricists disagree: the individual organism. The dispute is about what features of this system are present "independent" of the individual organism's learning, as part of its innate endowment, and what features are due to the organism's learning.

Unsurprisingly, most definitions of learning in psychology make it definitional that it is something that is done by an individual, so evolution would obviously not count as learning. For example, De Houwer, Barnes-Holmes, & Moors (2013)[18] define learning as "changes in the behavior of an organism that are the result of regularities in the environment *of that organism."*[19] 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. We could also stipulate that these changes are changes in the mental representations of organisms. (Just as the

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

[18] 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."

[19] De Houwer J, Barnes-Holmes D, Moors A. What is learning? On the nature and merits of a functional definition of learning. Psychon Bull Rev. 2013 Aug 1;20(4):631–42.

development of an immune response does not count as learning in psychology, the evolution of camouflage would not count as learning.) If so, we could give the following argument:

1. Evolution reliably creates cognitive architectures that are fitness-enhancing.
2. Cognitive architectures with more accurate representations are (all else equal) fitness enhancing.[20]
3. Evolution reliably creates cognitive architectures with more accurate representations of the environment.
4. If a process reliably creates cognitive architectures with more accurate representations of the environment, that process is (in a robust sense) learning.
5. Conclusion**:** Evolution is (in a robust sense) learning.

I think this argument suggests that at the right level of abstraction, one can view evolution as a learning process, without trivializing "learning." To be sure, this kind of "learning" doesn't fit many cognitive science nativists' definition of learning straightforwardly. For example, Carey says she thinks of learning as processes that involve treating data "as evidence." Certainly, evolution is not necessarily treating anything as evidence. On the other hand, 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 could be seen as a process of learning: of ensuring a better fit between our representations and the world. The "evidence" that evolutionary processes respond to is very coarse-grained;[21] it only tracks reproductive fitness. But in doing so, it can indirectly track the fitness of representations to the world.[22]

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 data, in order to better approximate some function. To take a classic and very 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.

In their classic (and notorious) *Perceptrons* (1968), Minsky and Paper 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.

[20] This has, of course, been challenged by people like Hoffman, Prakash, and Singh (2015). But also see formidable criticisms of Hoffman, like Cohen (2015)

[21] 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.

[22] 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.

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.[23]

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, 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 AGI 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, or the substantive issue of whether there is a route to AGI 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

[23]Such FP, Madhavan V, Conti E, Lehman J, Stanley KO, Clune J. 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

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—this favors practical nativism. The more efficient evolutionary techniques are, the better they are as an alternative—this favors 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.[24]

A 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.[25]

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

[24]Baldassarre G, Santucci VG, Cartoni E, Caligiore D. 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.

[25] As with so many foundational issues in AI, this point was anticipated by Turing 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."

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

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 has 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 'hones 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 nativism: arguments from safety

Another different kind of practicality concern has to do with the risks associated with the AGI that is created. A slogan for practical empiricism might be: "empiricist approaches are more powerful than the human-like nativist approach because they enable solutions that we could humans never dream of" (see: the argument from ambition, below). But the safety argument for nativism is that this very feature of empiricism makes it more dangerous. Imagine if empiricist approaches result in systems that solve the same problems that we do, but without utilizing our (partially innate) concepts of agency, objecthood, and causation. If our human value systems are specified partially in terms of these concepts, or our if ability to understand other agents depends on their sharing these concepts, then it might be impossible for us to control or understand the actions of an AGI that does not share these concepts. A route to AGI that begins with human-like innate machinery, the safety nativist says, will be more likely to produce AGIs that are comprehensible and value-aligned. Of course, nativist AIs might *still* be unsafe, because we did not specify the "right" innate machinery, or for other reasons. But the safety nativist argues that there is at least a greater chance of safe AI through nativist means, because empiricist approaches are likely to produce more inhuman AGI.

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 natural language processing engineer 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, obviously, 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 being 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? (Note that the argument from safety would reply that a system with "better" 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, Liz Spelke writes that "core knowledge captures fundamental properties of space, objects, and agency" (Spelke and Blass). On the other hand, if this is the case, then why not let systems learn these fundamental properties themselves? Whether this is possible depends on the strength of learnability arguments.

IV. Conclusion

In this paper, I have argued that nativism and empiricism are a fruitful lens for thinking about AI. Linking these debates to historical debates between nativists and empiricists helps us frame these debates as debates about the *general nature* of intelligence, not just human intelligence. 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?

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 AGI. 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. The question is likely only to be settled by progress in AI research. For now, we can expect that different researchers will make different bets on nativism and on empiricism, and a thousand flowers will bloom.

Works cited

Botvinick M, Barrett DG, Battaglia P, de Freitas N, Kumaran D, Leibo JZ, et al. Building machines that learn and think for themselves. Behavioral and Brain Sciences. 2017;40.

Buckner C. Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese. 2018 Dec 1;195(12):5339–72.

Carey S. Précis of "The Origin of Concepts." Behav Brain Sci. 2011 Jun;34(3):113-124-162.

Carey S. The Origin of Concepts. Oxford University Press; 2009. 609 p.

Chomsky N. Aspects of the theory of syntax. Cambridge, MA: MITPress. 1965.

Cohen J. Perceptual representation, veridicality, and the interface theory of perception. Psychon Bull Rev. 2015 Dec 1;22(6):1512–8.

De Houwer J, Barnes-Holmes D, Moors A. What is learning? On the nature and merits of a functional definition of learning. Psychon Bull Rev. 2013 Aug 1;20(4):631–42.

Domjan M. The principles of learning and behavior. Nelson Education; 2014.

Elman JL, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D, Plunkett K. Rethinking Innateness: A Connectionist Perspective on Development. Reprint edition. Cambridge, Mass.: A Bradford Book / The MIT Press; 1997. 447 p.

Hoffman DD, Singh M, Prakash C. The Interface Theory of Perception. Psychon Bull Rev. 2015 Dec 1;22(6):1480–506.

Hofstadter DR. Gödel, Escher, Bach: an eternal golden braid. New York: Basic Books; 1979.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building Machines That Learn and Think Like People. arXiv:160400289 [cs, stat] [Internet]. 2016 Apr 1; Available from: http://arxiv.org/abs/1604.00289

Levinovitz A. The mystery of Go, the ancient game that computers still can't win. Wired Magazine. 2014

Marcus G. Deep Learning: A Critical Appraisal. arXiv:180100631 [cs, stat] [Internet]. 2018 Jan 2; Available from: http://arxiv.org/abs/1801.00631

Marcus G. Innateness, AlphaZero, and Artificial Intelligence. arXiv:180105667 [cs] [Internet]. 2018 Jan 17; Available from: http://arxiv.org/abs/1801.05667

1.

Marcus G. The algebraic mind: Integrating connectionism and cognitive science. MIT press; 2001.

Margolis E, Laurence S. In defense of nativism. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition. 2013;165(2):693–718.

Minsky M, Papert S. Perceptrons, Reissue Of The 1988 Expanded Edition With A New Foreword By Léon Bottou. The MIT Press; 2017 .

Ramsey W, Stich S. Connectionism and Three Levels of Nativism. Synthese. 1990;82(2):177–205.

Samet J. Troubles with Fodor's nativism. Midwest Studies in Philosophy. 1987;10(1):575–594.

Samet J, Zaitchik D. Innateness and Contemporary Theories of Cognition. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy [Internet]. Fall 2017. Metaphysics Research Lab, Stanford University; 2017. Available from: https://plato.stanford.edu/archives/fall2017/entries/innateness-cognition/

Santoro A. Thoughts on "A Critique of Pure Learning," Zador (2019) [Internet]. Medium. 2019 [cited 2020 Jan 31]. Available from: https://medium.com/@adamsantoro/thoughts-on-a-critique-of-pure-learning-zador-2019-820a7dbbc783

Spelke ES, Kinzler KD. Innateness, Learning, and Rationality. Child Dev Perspect. 2009 Aug;3(2):96–8.

Spelke E. Initial knowledge: six suggestions. Cognition. 1994 Apr 1;50(1):431–45.

Spelke ES, Blass JA. Intelligent machines and human minds. Behavioral and Brain Sciences [Internet]. 2017 ed [cited 2018 Oct 30];40. Available from: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/intelligent-machines-and-human-minds/80A4A2D9E13C3E19651065CE0895FD65

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science. 2018 Dec 7;362(6419):1140–4.

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017;550(7676):354.

Skinner BF. Verbal behavior. New York: Appleton-Century-Crofts; 1957.

Such FP, Madhavan V, Conti E, Lehman J, Stanley KO, Clune J. 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

Sutton RS. The Bitter Lesson [Internet]. 2019 [cited 2020 Feb 2]. Available from: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.

Samet J. Troubles with Fodor's nativism. Midwest Studies in Philosophy. 1987;10(1):575–594.

Shulman C, Bostrom N. How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects. Journal of Consciousness Studies. 2012;19(7–8):7–8.

Turing AM. Computing machinery and intelligence. Mind. 1950 Oct 1;LIX(236):433–60.

Weber B (1996) Mean chess-playing machine tears at meaning of thought. New York Times, 19 February. Available at: www.rci.rutgers.edu/~cfs/472_html/Intro/NYT_Intro/ChessMatch/MeanChessPlaying.html (accessed 24 July 2011).

Zador AM. A critique of pure learning and what artificial neural networks can learn from animal brains. Nat Commun [Internet]. 2019 Aug 21;10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6704116/