Does AI Need Innate Conceptual Primitives?

Chaitanya K. Joshi

April 2019

This paper extends the Rationalism versus Empiricism debate to building Artificially Intelligent systems. Triggered by the explosion of work in learning-based approaches to AI, the debate centers around the need to embed innate conceptual primitives in AI systems akin to those in human beings; or whether such primitives can be *learnt* directly from interacting with the environment that the system is deployed in. I shall argue that human conceptual primitives themselves are a result of empirical learning over millions of years of evolution. Hence, building better learning mechanisms instead of handcrafting human primitives into AI systems is the most viable approach for building powerful AI, both from a technical and philosophical standpoint.

Background The debate of Rationalism versus Empiricism has been raging among philosophers of science for centuries. The rationalist camp, including Plato and Immanuel Kant, hold that there are innate concepts and knowledge built into humans at birth. On the other hand, empiricists such as John Locke and David Hume believe that humans start from a *blank slate* and acquire all our knowledge through experience and observation. The debate turned from philosophy to psychology over the past 50 years when scientists started asking questions about how cognition and learning develops in human beings. This debate is commonly titled, Nature versus Nurture: cognition emerging from innate and natural primitives versus an emphasis on learning from the environment. Today, the same debate rages on in the Artificial Intelligence community in a new form. What should the focus of building AI systems be: introducing conceptual primitives into models through handcrafted rules versus building learning mechanisms which automatically infer the primitives from observations?

Before we begin, it is important to highlight a demarcation between building *Narrow AI*, where one is trying to solve specific tasks in limited domains, and *Broad/General AI*, where the focus in on building systems which display human-like intelligence. General AI is vaguely defined: only when we can build strong narrow AIs in several domains can we start building towards general AI. Thus, the rationalism versus empiricism debate is fundamental for determining what direction future research on narrow AI (and by extension general AI) should take. This paper argues for focusing on the empiricist approach, also known as *machine learning*.

In the following paragraphs, I shall introduce *deep learning*, a sub-field of machine learning that powers the best AI systems today. I shall then critically analyze two major philosophical criticisms of deep learning from the rationalist camp: the lack of generally applicable conceptual primitives, and the brittle bond with datasets. Finally, I shall present technical results from the AI community that refute or accommodate the rationalists' arguments. I shall use these results to argue that deep learning is the most viable approach for building AI today, both from a technical and philosophical standpoint.

Deep Learning Till around 2010, most successful AI systems for complex tasks, such as the processing of natural languages or computer images, were rationalist systems: their success depended on handcrafting domain-specific rules and complex mathematical models of how a system should behave in an environment. In recent years, fueled by rapidly growing computational power and availability of high quality datasets, there has been an explosion of work successfully applying machine learning to domains such as languages and images. The new wave of narrow AI systems today is powered by deep learning, where the emphasis is on starting from relatively simple models, and learning compositional concepts directly from the environment. Deep learning models are loosely inspired by neurons connected to each other in the human brain, and are termed as *artificial neural networks*. The connections between several thousands of neurons in a neural network are randomly initialized and subsequently updated during the model's training process to reflect an understanding of the environment. Neural networks can be used to make inferences and predictions by providing inputs which are propagated and processed by a chain of neurons in the network.

Lack of Generally Applicable Conceptual Primitives Critics of deep learning, the most prominent among which is NYU psychologist Gary Marcus, point out that neural networks are *black-boxes*. Even though they might outperform handcrafted models for various tasks, we will never know *why* they do so because they do not provide any innate conceptual primitives or mechanisms for their predictions. Marcus laments that AI has been leaning towards machine learning, arguing human-grade intelligence and reasoning can not emerge without basic conceptual primitives [3]. He points to research in evolutionary psychology which shows that humans, even newborn infants, are endowed with the basic ability to perceive persons, sets and places [10, 4]. He further highlights that the brain has structural parts that are responsible for very specific tasks such as vision or learning languages [6]. Although we do not fully understand how they function, there is ample evidence from biology that they exist. Thus, in the rationalist view, building a system for any task would first involve carefully understanding and interpreting the primitives involved in solving it.

My response to this argument is to question the innateness of the conceptual primitives themselves. It can be argued that innate concepts emerge from observations: Albert Einstein famously said that concepts (not limited to concepts in Physics) are summaries of repeated experiences. Thus, attempting to handcraft *all* possible primitives for complex tasks would be an inefficient strategy for building AI systems, simply due to the diversity of experiences possible. As a solution, pioneers of deep learning, such as Juergen Schmidhuber of the University of Lugano and Yann LeCun of Facebook AI, have independently proposed to build models that continuously form and evolve their own conceptual primitives about their environment, solely through interaction within the environment [8, 9].

A uniting feature of human primitives and those formed by neural networks is that both are extremely effective for cognition and reasoning (as pointed out by Marcus himself, in the case of humans). A recent example of the success of learnt primitives in AI systerms is the work of Ha and Schmidhuber [2], wherein they implement the idea of learning representations of an environment for playing video games. They used a neural network to represent the most important aspects of the game engine, which the network then used to train itself to play the game without explicitly operating in the game environment. This is akin to 'dreaming' about playing the game: how a professional gamer might think of strategies while not directly engaged in the act of playing. When the model was deployed in the actual game environment, they found that the strategies learnt through dreaming translated very well to the game. [1]

[1] Summary video: `https://www.youtube.com/watch?v=gvjCu7zszbQ`

Furthermore, research in evolutionary biology tells us that the innate conceptual primitives in the human brain are a product of evolution. Continuing with the theme of video games, work by psychologists at UC Berkeley [1] found that explicitly removing primitives that humans have built up about video games drastically decreases human performance on games. Interestingly, deep learning models had no degradation in performance as the video game environments transformed in their experiments. This dichotomy highlights that conceptual primitives emerge through a similar learning process in both humans and neural networks. [2] [3]

Brittle Bond with Datasets Marcus further argues that AI systems built through a principled study of conceptual primitives involved in tasks are able to generalize to variations in the tasks. On the other hand, learning-based systems are brittle-ly tied to the datasets used for their training. If a neural network is trained to identify breeds for dogs $b_1, b_2, ..., b_n$ from dataset D_1, it will generalize poorly to a new dataset D_2 which might contain additional breeds of dogs $b_{n+1}, b_{n+2}, ..., b_{n+m}$, previously un-encountered by the model in $D1$. This is a repackaging of Hume's problem of induction or the curve fitting problem in data analysis. The usual solution to this problem is to assume the Principle of Uniformity of Nature and say that the data in the future will come from the same probability distribution as the data available today.

Marcus extends this argument by citing the work of Noam Chomsky, a linguist who famously said that real-world tasks such as language understanding and reasoning are far too complex for representing them as probability distributions. In response, I would like to point out that neural networks, given large amounts of compute and data, are able to learn meaningful statistical representations of complex tasks, such as those involving language. Recent work on language understanding at OpenAI, for example, focuses on large-scale learning of the structure of English through word prediction in partially complete sentences [5, 7]. Through experiments where the trained model was used to generate new sentences, they showed that the neural network had a command over the grammatical structure of English that is clearly superior to most people. Their approach outperformed humans by large margins at several benchmark language generation and understanding tasks. [4]

It is easy to criticise OpenAI's lengthy and expensive training process (equivalent to hundreds of GBs of text and thousands of human hours). After all, human children learn the structure of languages at a very rapid rate. However, one may draw a parallel between the initial training of a neural network and the long process of evolution that has lead to the primitives enabling the children of today to learn faster. Recent research has highlighted the effectiveness of taking 'well-evolved' neural networks (such as OpenAI's language model) and re-training or *finetuning* them on new tasks. If the conceptual primitives learnt by the base model are powerful, this approach leads to very fast learning and good performance for the new task or environment.

In conclusion, this paper argues that future research in AI should focus on improving learning mechanisms instead of attempting to embed human conceptual primitives into narrow AI systems. When thinking about building Artificial General Intelligence from narrow AI systems, we should take inspiration from evolution in a similar manner as airplanes are inspired by birds: Primitives in the human brain that enable us to learn languages or concepts so effectively are a product of evolution. Thus, scaling up powerful learning mechanisms for AI systems using more computation [11] and better data [12] is the most effective way today to learn conceptual primitives about any new task.

[2] Summary video: `https://www.youtube.com/watch?v=Ol0-c9OE3VQ`

[3] You can play this game yourself: `https://rach0012.github.io/humanRL_website/`

[4] Summary video: `https://www.youtube.com/watch?v=8ypnLjwpzK8`

References

[1] Rachit Dubey, Pulkit Agrawal, Deepak Pathak, Thomas L. Griffiths, and Alexei A. Efros. Investigating human priors for playing video games. In *ICML*, 2018.

[2] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pages 2450–2462, 2018.

[3] Gary F Marcus. *The algebraic mind*. The MIT Press, 2001.

[4] Gary F Marcus. *The birth of the mind: How a tiny number of genes creates the complexities of human thought*. Basic Civitas Books, 2004.

[5] OpenAI. Better language models and their implications. `https://openai.com/blog/better-language-models/`. Accessed: 2019-04-23.

[6] Steven Pinker. The language instinct (1994/2007). 2007.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

[8] Juergen Schmidhuber. One big net for everything. *arXiv preprint arXiv:1802.08864*, 2018.

[9] Tom Simonite. Teaching machines to understand video could be the key to giving them common sense. `https://www.technologyreview.com/s/603803/facebooks-ai-chief-machines-could-learn-common-sense-from-video/`. Accessed: 2019-04-23.

[10] Elizabeth Spelke. Initial knowledge: Six suggestions. *Cognition*, 50(1-3):431–445, 1994.

[11] Richard Sutton. The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`. Accessed: 2019-04-23.

[12] Max Welling. Do we still need models or just more data and compute? `https://staff.fnwi.uva.nl/m.welling/wp-content/uploads/Model-versus-Data-AI-1.pdf`. Accessed: 2019-04-23.