What have we learned about artificial intelligence from studying the brain?

Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University
Center for Brains, Minds, and Machines, MIT

April 3, 2023

Abstract

Neuroscience and artificial intelligence (AI) share a long, intertwined history. It has been argued that discoveries in neuroscience were (and continue to be) instrumental in driving the development of new AI technology. Scrutinizing these historical claims yields a more nuanced story, where AI researchers were loosely inspired by the brain, but ideas flowed mostly in the other direction.

The current excitement

We now have artificial intelligence (AI) systems that can converse with us, beat us at our own games, and help us solve scientific problems like protein folding and fusion reactor design. It is significant that these systems achieve human-level profiency using machinery that is inspired by the human brain. The idea that neural networks are not only similar to the brain, but are successful precisely because of this similarity, has generated considerable excitement about the possibility that studying the brain will unlock the recipe for general intelligence [1, 2, 3]. For example, Hassabis and colleagues [1] assert that "better understanding biological brains could play a vital role in building intelligent machines." Similarly, Macpherson and colleagues [2] write: "Advances in neuroscience... have given rise to a new generation of in silico neural networks inspired by the architecture of the brain." Zador and colleagues [3] make an even stronger assertion: "Neuroscience has long been an essential driver of progress in artificial intelligence (AI). We propose that to accelerate progress in AI, we must invest in fundamental research in NeuroAI."

Indeed, considerable resources have already been mobilized to seek biological inspiration for AI. The company DeepMind was founded on the principle that engineering intelligent systems and understanding the brain are part of a single project. Other companies, such as Vicarious and Numenta, follow similar founding principles. Established companies such as Intel and IBM have invested in neuromorphic computing. The federal government of the United States has initiated numerous funding programs directed at the intersection of AI and neuroscience. Major philanthropic organizations have created centers and institutes dedicated to the same objective.

In light of these efforts, it's worth asking: what have we learned about AI from studying the brain?

The innocent eye

Most AI researchers would say that they are looking for computational principles derived from biology, rather than particular details at the level of anatomical organization or biochemistry. This sounds appealing, but it runs into conceptual difficulties. How do we derive computational principles from biology? Principles are not resting on the surface of measurement data, waiting to be observed; there is no *innocent eye* that can "just look" at the data (see [4] for more discussion). Even if we allow that some data analysis has to intervene between measurement and interpretation, this often simply replaces the innocent eye with an innocent algorithm, the output of which must then be interpreted. The limiting factor is not our ability to extract structure from data, but rather our ability to specify what kind of structure we are looking for in the first place. This, in turn, depends on our theoretical arsenal *prior* to observing the data. Where does this come from?

Many of the most impactful ideas in neurobiology have come from other fields. Shannon's information theory inspired efficient coding models; Fourier analysis inspired models of spatial frequency analysis in the visual system; statistical mechanics inspired attractor network models of memory; statistical decision theory inspired evidence accumulation models of perceptual decision making; the list goes on and on. In all of these cases, it was not the biologists left to their own devices who invented new technical concepts. It was not the case that biologists "just looked" at the firing of neurons and then invented information theory, Fourier analysis, etc. The theoretical arsenal came from elsewhere, invented independently of discoveries in neurobiology.

Lessons from history

When people talk about biologically inspired AI, they often refer to a few canonical examples. One is the foundational work by McCulloch and Pitts [5], and later by Rosenblatt [6], which showed that neural networks, loosely inspired by biological neurons, were capable of logical computation and pattern recognition. A second is the convolutional neural network (convnet), inspired by the organization of visual cortex. A third is reinforcement learning, inspired by studies of animal learning. These examples deserve careful scrutiny. In the interest of brevity, I will focus on the second and third examples, because these highlight the specific intellectual contributions of neuroscience to AI which go beyond general ideas about neural computation.

Convolutional neural networks

In 1980, Fukushima published a seminal paper introducing his neocognitron architecture [7], which was based on the single-unit recordings of visual cortex reported by Hubel and Wiesel [8, 9, 10]. The first practical convnet (trained using backpropagation) was developed by LeCun and colleagues [11], which they applied to handrwritten digit classification. With advances in computing power and data set size, convnets came to dominate computer vision [12]. They subsequently fed back into neuroscience, driving new experimental and theoretical work [13]. Thus, the history of convnets seems like a paradigmatic case study of positive feedback between neuroscience and AI.

To substantiate this claim, we need to look more closely at what Hubel and Wiesel actually found. In their 1959 paper, they reported the existence of "simple cells' in primary (striate) visual cortex, which respond selectively to spots of light on the retina. Simple cells have retinotopic

receptive fields, responding strongly to light in particular locations on the retina. The receptive fields also typically have an inhibitory region flanking the excitatory region (or vice versa). Hubel and Wiesel noted a number of variations across simple cells:

> Some fields had long narrow central regions with extensive flanking areas: others had a large central area and concentrated slit-shaped flanks. In many fields the two flanking regions were asymmetrical, differing in size and shape; in these a given spot gave unequal responses in symmetrically corresponding regions. In some units only two regions could be found, one excitatory and the other inhibitory, lying side by side. ([8], pp. 579-580)

These variations are significant because a critical feature of the neocognitron, and virtually all subsequent convnets, is the assumption that the receptive fields of cells within a convolutional layer are shifted copies of one another. Another form of variation reported by Hubel and Wiesel was in the size of receptive fields, ranging from 4° to 10°. This is a substantial range of variation when one considers that the size of foveal vision (the high acuity region of the visual field) is 1°. Again, this directly contradicts the assumption of shift invariance.

Hubel and Wiesel reported other properties of simple cells that were not incorporated into the neocognitron or its descendants: baseline firing rate, motion selectivity, and ocular selectivity. These properties also varied across cells. For example:

> Thirty-six units in this study could be driven only from one eye, fifteen from the eye ipsilateral to the hemisphere in which the unit was situated, and twenty-one from the contralateral. Nine, however, could be driven from the two eyes independently. Some of these cells could be activated just as well from either eye, but often the two eyes were not equally effective, and different degrees of dominance of one eye over the other were seen. ([8], p. 584)

It should be clear by now that the assumption of shift invariance was biologically questionable, just on the basis of this one study of simple cells. Moreover, a later study by Hubel and Wiesel [14], not cited by Fukushima, showed that receptive field size increases with eccentricity away from the fovea—another violation of shift invariance.[1]

In summary, convnets were undoubtedly inspired by studies of the visual system, but from the beginning they made assumptions that directly contradicted the biological data. Those biologically implausible assumptions (particularly shift invariance) turned out to be of great practical importance, because it meant that weights could be shared by convolutional filters, dramatically reducing the number of parameters that needed to be learned.

Reinforcement learning

The theory of reinforcement learning coalesced in the 1980s thanks to the work of Sutton and Barto [17, 18, 19], who formalized the structure of the problem that needed to be solved and developed algorithms to solve it—notably the temporal difference, or TD, learning algorithm. A wide variety of similar algorithms had previously been applied to reinforcement learning problems with some success, but it was not clear up to that point why they worked (or didn't work). The situation changed dramatically once the logic of TD learning algorithms was understood, leading to many

[1] Some recent work in computer vision has begun to incorporate eccentricity dependence into convnets [15, 16].

generalizations and improvements. These algorithms continue to be the workhorses of modern reinforcement learning systems [20] (though the history of reinforcement learning is much richer than TD; see [21]).

Sutton and Barto were remarkable for another reason: they had an unusually interdisciplinary view of the subject, drawing upon ideas from psychology and neuroscience. They wrote a number of papers showing how their algorithms were accurate models of classical conditioning phenomena, addressing some of the problems that vexed earlier models. Although the detailed biological data on the neural mechanisms of classical conditioning was not yet available, Sutton and Barto were aware of developments in neuroscience (e.g., Kandel's studies of habituation) and considered the biological plausibility of their learning rules.

The question here is whether the biological and behavioral data directly inspired the development of TD learning algorithms. To answer this question, it's useful to examine the progression of ideas from their first major paper on classical conditioning [22] to the book chapter published a decade later [23]. The two publications were based on largely the same body of empirical data. A key difference is that the second publication invoked the TD learning algorithm, the logic of which had been worked out a few years earlier. If the empirical data were a truly powerful source of inspiration, then one might have expected that the TD learning algorithm would already have been invented in 1981, when Sutton and Barto were first thinking about classical conditioning. Instead, what happened in the intervening years was a slow process of clarifying the structure of the reinforcement learning problem, which eventually fed back into the models of classical conditioning.

In summary, the TD algorithm was undoubtedly inspired by studies of animal learning, but only in a fairly weak sense. Sutton and Barto were interested in explaning how animals learn and also how to build machines that learn. It turned out that doing the latter was useful for doing the former. It was only *after* the core engineering problem had been solved that the appropriate computational framework for animal learning came into view. It's also worth noting that biology played very little role in this story; all of the exciting biology (dopamine, the basal ganglia, etc.) came later.

Convergence

Instead of looking for inspiration, a more plausible (but weaker) heuristic is to look for convergence: If engineers propose an algorithm and neuroscientists find evidence for it in the brain, it is a pretty good clue that the algorithm is on the right track (at least from the perspective of building human-like intelligence). This convergence heuristic is consistent with current computational neuroscience practice, where AI has historically provided a fund of ideas for biological theories. It is also consistent with current AI practice, where researchers are primarily looking for directional signals from neuroscience (is this roughly what the brain does?) rather than specific algorithms.

The strongest constraints on algorithms will always come from the structure of the problems that need to be solved, since engineers are paid to solve those problems rather than explain how the brain works. Happily, algorithms optimized for solving engineering problems frequently turn out to be successful models of brain function. This is a reason for optimism about future synergies between AI and biology.

Acknowledgments

I'm grateful to Andy Barto, Venki Murthy, Chris Summerfield, Gabriel Kreiman, Chris Bates, and Jay Hennig for comments on an earlier draft. This work was supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF1231216.

References

[1] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95:245–258, 2017.

[2] Tom Macpherson, Anne Churchland, Terry Sejnowski, James DiCarlo, Yukiyasu Kamitani, Hidehiko Takahashi, and Takatoshi Hikida. Natural and artificial intelligence: A brief introduction to the interplay between ai and neuroscience research. *Neural Networks*, 144:603–613, 2021.

[3] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nature Communications*, 14:1597, 2023.

[4] Samuel J Gershman. Just looking: The innocent eye in neuroscience. *Neuron*, 109:2220–2223, 2021.

[5] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[6] F Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[7] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[8] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148:574–591, 1959.

[9] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160:106–154, 1962.

[10] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28:229–289, 1965.

[11] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[13] Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 33:2017–2031, 2021.

[14] David H Hubel and Torsten N Wiesel. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *Journal of Comparative Neurology*, 158:295–305, 1974.

[15] Francis X Chen, Gemma Roig, Leyla Isik, Xavier Boix, and Tomaso Poggio. Eccentricity dependent deep neural networks: Modeling invariance in human vision. *AAAI Spring Symposium Series*, 2017.

[16] Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

[17] Richard S Sutton. Single channel theory: A neuronal theory of learning. *Brain Theory Newsletter*, 4:72–75, 1978.

[18] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on Systems, Man, and Cybernetics*, pages 834–846, 1983.

[19] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[21] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[22] Richard S Sutton and Andrew G Barto. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88:135–170, 1981.

[23] Richard S Sutton and Andrew G Barto. Time-derivative models of Pavlovian reinforcement. *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, 1990.