Biological Cybernetics

PERSPECTIVE

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

Samuel J. Gershman[1,2]

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 proficiency 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 (Hassabis et al. 2017; Macpherson et al. 2021; Zador et al. 2023). For example, Hassabis et al. (2017) assert that "better understanding biological brains could play a vital role in building intelligent machines." Similarly, Macpherson et al. (2021) write: "Advances in neuroscience... have given rise to a new generation of in silico neural networks inspired by the architecture of the brain." Zador et al. (2023) 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 USA has initiated numerous funding programs directed at the intersection of AI and neuroscience. Major philanthropic organizations have created centers, institutes, and conferences dedicated to the same objective (e.g., the Kempner Institute at Harvard University, the Center for Brains, Minds, and Machines at MIT, and the NeuroAI program at Cold Spring Harbor Laboratory).

In light of these efforts, it is 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 *inno-*

✉ Samuel J. Gershman
gershman@fas.harvard.edu

1   Department of Psychology and Center for Brain Science, Harvard University, Cambridge, USA, Cambridge, USA

2   Center for Brains, Minds, and Machines, MIT, Cambridge, USA

cent eye that can "just look" at the data (see Gershman 2021 for more discussion).[1] 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.

One response to this argument is that there is no reason to expect it could be any other way—biological inspiration will always be "loose" because it is infeasible (and probably misguided) to transfer all the details. I agree with this point, but the problem is that it threatens to render the notion of inspiration vacuous. Even if we were able to state some non-vacuous version, there is the remaining challenge of identifying which details should be transferred. I submit that it is impossible to do so without already coming to the data with a computational framework.

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 (1943), and later by Rosenblatt (1958), which showed that neural networks, loosely inspired by biological neurons, were capable of logical computation and pattern recognition. A second is the massive parallelism of neural computation, which inspired "cooperative" algorithms for parallel constraint satisfaction (Grimson 1981; Ackley et al. 1985). A third is the convolutional neural network (convnet), inspired by the organization of visual cortex. A fourth is reinforcement learning, inspired by studies of animal learning. These examples deserve careful scrutiny. In the interest of brevity, I will focus on the third and fourth

---

[1] In Gershman (2021), it is also argued that there is no "innocent algorithm" for analyzing data without making certain assumptions.

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 (Fukushima 1980), which was based on the single-unit recordings of visual cortex reported by Hubel and Wiesel (1959, 1962, 1965). The first practical convnet (trained using backpropagation) was developed by LeCun et al. (1989), which they applied to handwritten digit classification. With advances in computing power and data set size, convnets came to dominate computer vision (Krizhevsky et al. 2017). They subsequently fed back into neuroscience, driving new experimental and theoretical work (Lindsay 2021). 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. (Hubel and Wiesel 1959, 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. (Hubel and Wiesel 1959, 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 (1974), not cited by Fukushima, showed that receptive field size increases with eccentricity away from the fovea—another violation of shift invariance.[2]

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.

While biological implausibility might be an asset for AI, one could also argue that the complexity of the visual system is a source of untapped computational potential. It has long been known that models based on Hubel and Wiesel's characterization of visual cortex explain a surprisingly small fraction of the variance in neural activity (Olshausen and Field 2006). More recent models, despite their sophistication, still do only slightly better than an untrained model, and remain considerably below the noise ceiling (Zhuang et al. 2021). Thus, early sensory areas might still harbor interesting secrets. Unlocking these secrets will require more than just looking at activity in visual cortex, since if that were the case then we'd have unlocked them long ago. Instead, we should look toward AI as a source of ideas. The study by Zhuang et al. (2021), even if it is not the last word on visual cortex, exemplifies the way in which models from AI have been instrumental in driving progress in visual neuroscience.

Reinforcement learning

The theory of reinforcement learning coalesced in the 1980s thanks to the work of Sutton (1978), Barto et al. (1983), and Sutton (1988), 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 generalizations and improvements. These algorithms continue to be the workhorses of modern reinforcement learning systems (Mnih et al. 2015) (though the history of reinforcement learning is much richer than TD; see Sutton and Barto 2018).

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 were 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 is useful to examine the progression of ideas from their first major paper on classical conditioning (Sutton and Barto 1981) to the book chapter published a decade later (Sutton and Barto 1990). 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 explaining 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 is also worth noting that biology played very little role in this story; all of the exciting biology (dopamine, the basal ganglia, etc.) came later (Houk et al. 1995; Schultz et al. 1997).

---

[2] Some recent work in computer vision has begun to incorporate eccentricity dependence into convnets (Chen et al. 2017; Deza and Konkle 2020).

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). While convergence is neither necessary nor sufficient for ensuring the quality of an algorithm, it can nonetheless provide useful guidance when (as is the case now) existing algorithms are not well separated based on current AI benchmarks. In effect, neuroscience can act as another benchmark (Schrimpf et al. 2020), though we should avoid reducing the complexity of the brain to a single number that can be optimized (cf. Goodhart's law).

The 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.

A good recent example of convergence is the study of stochastic computation. It has long been known that neural activity appears stochastic at multiple levels (Faisal et al. 2008). Researchers have speculated about the computational function of this stochasticity: escaping from local optima, sampling probability distributions, exploration, and regularization, to name a few. These speculations are typically grounded in engineering ideas. For example, the sampling hypothesis comes from Monte Carlo approximation techniques (Buesing et al. 2011; Gershman et al. 2012), and the exploration hypothesis comes from reinforcement learning theory (Gershman and Ölveczky 2020). Researchers have also been interested in stochasticity from an energy efficiency perspective: stochastic spike-based neuromorphic chips now exist that achieve dramatically lower energy demands (Roy et al. 2019). This line of work is notable for its commitment to biological plausibility, though it is not yet clear which biological details matter for performance. A comparable level of commitment to biological plausibility has not yet penetrated modern AI at large, due to the fact that there are many non-biological options that are more convenient and effective (though usually more energy-intensive). This further highlights the fact that biology is not necessary for progress in AI, but it can serve as a useful directional signal for certain desiderata like energy efficiency.[3]

---

[3] When the Kempner Institute was created at Harvard, I suggested to the directors that if they really wanted to advance biologically inspired AI, they should restrict the compute budget to the wattage of a light bulb, which is all the brain needs. My suggestion was not followed.

Conclusion

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.

Author Contributions S.G. wrote the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cogn Sci 9:147–169

Barto AG, Sutton RS, Anderson CW (1983) Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern 834–846

Buesing L, Bill J, Nessler B, Maass W (2011) Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Comput Biol 7:e1002211

Chen FX, Gemma R, Leyla I, Xavier B, Tomaso P (2017) Eccentricity dependent deep neural networks: modeling invariance in human vision. In: AAAI spring symposium series

Deza A, Konkle T (2020) Emergent properties of foveated perceptual systems. arXiv:2006.07991

Faisal AA, Selen LPJ, Wolpert DM (2008) Noise in the nervous system. Nat Rev Neurosci 9:292–303

Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36:193–202

Gershman SJ (2021) Just looking: the innocent eye in neuroscience. Neuron 109:2220–2223

Gershman SJ, Ölveczky BP (2020) The neurobiology of deep reinforcement learning. Curr Biol 30:R629–R632

Gershman SJ, Vul E, Tenenbaum JB (2012) Multistability and perceptual inference. Neural Comput 24:1–24

Grimson WEL (1981) From images to surfaces: a computational study of the human early visual system. MIT Press, New York

Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-inspired artificial intelligence. Neuron 95:245–258

Houk JC, Davis JL, Beiser DG (1995) Models of information processing in the Basal Ganglia. MIT Press, New York

Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574–591

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154

Hubel DH, Wiesel TN (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol 28:229–289

Hubel DH, Wiesel TN (1974) Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. J Compar Neurol 158:295–305

Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1:541–551

Lindsay GW (2021) Convolutional neural networks as a model of the visual system: past, present, and future. J Cogn Neurosci 33:2017–2031

Macpherson T, Churchland A, Sejnowski T, DiCarlo J, Kamitani Y, Takahashi H, Hikida T (2021) Natural and artificial intelligence: a brief introduction to the interplay between ai and neuroscience research. Neural Netw 144:603–613

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518:529–533

Olshausen BA, Field DJ (2006) van Hemmen L, Sejnowski T (eds) What is the other 85 percent of V1 doing, vol 23, pp 182–211

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386–408

Roy K, Jaiswal A, Panda P (2019) Towards spike-based machine intelligence with neuromorphic computing. Nature 575:607–617

Schrimpf M, Kubilius J, Lee MJ, Murty NAR, Ajemian R, DiCarlo JJ (2020) Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron 108:413–423

Schultz W, Dayan P, Read Montague P (1997) A neural substrate of prediction and reward. Science 275:1593–1599

Sutton RS (1978) Single channel theory: a neuronal theory of learning. Brain Theory Newsl 4:72–75

Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3:9–44

Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88:135–170

Sutton RS, Barto AG (1990) Time-derivative models of Pavlovian reinforcement. In: Learning and computational neuroscience: foundations of adaptive networks

Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT Press, New York

Zador A, Escola S, Richards B, Ölveczky B, Bengio Y, Boahen K, Botvinick M, Chklovskii D, Churchland A, Clopath C et al (2023) Catalyzing next-generation artificial intelligence through NeuroAI. Nat Commun 14:1597

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, Yamins DLK (2021) Unsupervised neural network models of the ventral visual stream. Proc Natl Acad Sci 118:e2014196118