Artificial intelligence versus biological intelligence: A historical overview

Roy de Kleijn

Leiden University

Published as de Kleijn, R. (2022). Artificial intelligence versus biological intelligence: A historical overview. In B. Custers & E. Fosch Villaronga (Eds.), *Law and artificial intelligence: Regulating AI and applying AI in legal practice.* Springer.

Abstract

The discipline of artificial intelligence originally aimed to replicate human-level intelligence in a machine. It could be argued that the best way to replicate the behavior of a system is to emulate the mechanisms producing this behavior. But what mechanisms exactly should we replicate, the brain or the cognitive faculties more directly? Early symbol-based AI systems paid little regard to neuroscience and were rather successful. However, since the 1980s, artificial neural networks has become a powerful AI technique that shows remarkable resemblance to what we know about the human brain. In this chapter, we will highlight some of the similarities and differences between artificial and human intelligence, the history of their interconnection, what they both excel at, and what the future may hold for artificial general intelligence.

1 Introduction

Humans are still considered to be smarter than machines, despite impressive progress in the field of artificial intelligence, with computers outperforming humans on several benchmark tasks. The big question that unites the fields of artificial intelligence and cognitive science remains unsolved: what makes humans so intelligent? It is surely not unreasonable to think that whatever it is that makes humans intelligent could be implemented in a computer system to make it as intelligent as a human. Unfortunately, we do not have a complete understanding yet of the origins of human intelligence, but we do have some idea based on both empirical evidence as well as some educated guesses.

First of all, we know that our cognitive faculties are physically implemented in the brain. This is a cantaloupe-sized mass of 1300 grams consisting of 10^{11} nerve cells called neurons, which are interconnected by around 10^{15} synapses. Neurons can transmit signals to other neurons, which can then send that signal to other neurons, et cetera, generating interneuronal communication. Information can then propagate through this network of interconnected neurons.[1] Simplified, a neuron receives a signal from other neurons, integrates these signals, and sends a signal when the sum of its received signals reaches a certain threshold. Depending on the strength of the synapse, such a signal can excite or inhibit other neurons. Computationally, what neurons do can be seen as a floating-point operation[2], of which the human brain can carry out 10^{18} per second.[3] The highest performing modern (as of 2021) personal computers can perform around 10^{14} floating-point operations per second, which makes them four magnitudes of order slower than a human brain. As such, whatever the brain *computationally* does in one minute could be performed on such a computer in a week. But, impressive as this may be, is it simply the speed of

1 Although it should be noted that it is unknown how anything but the most trivial information (e.g. stimulus intensity) is represented by neurons.

2 Simplified, a floating-point operation (FLOP) is an arithmetic operation (such as addition or multiplication) involving two real numbers. This is arguably what neurons and synapses do as well: multiplying incoming activation with the strength of the synapse.

3 Every synapse carrying out floating-point operations at 1000 Hz; McClelland 2009, Zador 2019.

floating-point operations that enables the human brain to produce intelligence? The fact that we do not actually know *how* to simulate a minute's worth of full human cognition, even given all the computing power in the world, suggests that we may need to understand more than mere floating-point operations in order to understand human intelligence.

But, even if we start with them, floating-point operations do not exist in a vacuum, they require *operands.* What exactly are the operands of the computations that produce the human mind? On a neural level they may be the activation values of neurons, but the relationship between neural activity and the mind is everything but clear. To create intelligent systems, do we need to recreate the brain, the mind, or neither? Several different schools of thought have dominated the field of artificial intelligence and cognitive science over the past century, but for the purpose of this chapter we will divide them roughly into symbolic and subsymbolic approaches of intelligence.

2 The beginnings of artificial intelligence and cognitive science

The origins of modern artificial intelligence cannot be seen separate from the origins of the field of cognitive science, which emerged from the ashes of the then-dominant psychological school of behaviorism.[4] In what is now known as the *cognitive revolution,* emphasis shifted from studying behavior to studying the computations producing it.[5] Moving beyond stimulus–response associations, concepts like reasoning and representations became the topic of study for the new fields of cognitive science and cognitive psychology. This new generation of researchers used models of mental processes to study the mind and behavior. Around the same time, a group of researchers interested in the idea of implementing intelligence in machines organized what is now known as the birthplace of the field of artificial intelligence: the Dartmouth Conference of 1956.[6] Although many topics were discussed in this two-month(!) meeting, one of the most direct outcomes of the Dartmouth Conference was the rise of symbolic artificial intelligence.

3 Symbolic AI and physical symbol systems

Allen Newell and Herbert Simon, two cognitive scientists who participated in the Dartmouth Conference, suggested that human intelligence is essentially symbol manipulation. And if humans are intelligent by virtue of their symbolic representation and manipulation, it could perhaps be possible to endow computer systems with this same capability. This position is now known as the *physical symbol system hypothesis* and was strengthened by the success of their computer programs *Logic Theorist*[7] and *General Problem Solver.*[8] Using manipulation of high-level symbols, the first could reason, and generate proofs for mathematical theorems, even improving on some proofs found by humans, while the second was a more general program to solve logical problems.

Newell and Simon's General Problem Solver would use means-end analysis to solve problems similar to how humans were thought to solve them, a paradigm now known as *reasoning as search.* How exactly humans solve unfamiliar problems was not well known, but Newell and Simon hypothesized that means-end analysis would be involved. Accordingly, they implemented this assumed human problem-solving technique into a computer system. Given a well-defined problem, it would cast it in terms of an initial state, a goal state, and operators that define the transition between two states. It would then solve the problem

4 At least, then-dominant in the United States. Behaviorism posits that psychology should limit itself to observing and predicting *behavior,* in contrast to *convictions* and *beliefs,* as only behavior could be a source of objective evidence. This behavior was regarded as the learned product of the interactions between an organism and its environment.

5 It could be argued that Chomsky's (1959) *Review of B. F. Skinner's Verbal Behavior* kickstarted the cognitive revolution. In this critique, Chomsky argued against the concept of language as purely learned behavior. For example, children are able to understand sentences they have never been exposed to before.

6 Attendees included Marvin Minsky, Claude Shannon, Allen Newell, Herbert Simon, W. Ross Ashby, and Ray Solomonoff, researchers who would become known as the founders of the field.

7 Newell et al 1957.

8 Newell and Simon 1961.

using heuristic search, narrowing the search space to make the search tractable.[9] First, it would evaluate the difference between the current state and the goal state. Second, it would find a transformation leading to a subgoal that reduces the difference between the current state and the goal state. It would then check if the transformation can be applied to the current state, and if not, would find another transformation. By iteratively transforming the symbolic representation of the initial state, the program could then find a solution to the problem. Newell and Simon demonstrated that their program could solve the *Towers of Hanoi* and *missionaries and cannibals* problems,[10] although it could be applied to any well-defined problem. Although immensely influential in cognitive psychology, some argue that the General Problem Solver failed as a psychological theory or as an explanation of human problem solving as the idea that problem solving relies on a general mechanism does not seem to hold.[11] Concepts from the General Problem Solver have now been incorporated in the more general Soar cognitive architecture.[12]

In the 1970s, progress was made in the field of *expert systems.* These systems attempted to implement the knowledge and decision making of human experts. In line with the physical symbol system hypothesis, knowledge would be represented symbolically in a *knowledge base* and reasoned with using an *inference engine.* Perhaps the most well-known example is MYCIN, an expert system developed in the early 1970s at Stanford. With a knowledge base of around 600 rules, it was designed to diagnose blood infections using a physician as an intermediate. The physician would be presented with a series of questions[13] and the system would then produce a list of possible diagnoses with certainty factors. Although the quality of MYCIN's prescribed antimicrobials was as least as good as human faculty specialists at Stanford,[14] the system was never used in medical practice. Ethical and legal issues surrounding liability and reliability of such a novel technique in medical practice, most of which have not been solved as of yet, detracted from its usefulness.

To develop an expert system, knowledge from human experts needs to be extracted and represented in its knowledge base and inference engine, a process known as *knowledge acquisition.* In other words, it requires the transfer of symbolic knowledge from the human mind to an artificial system. In its earliest form, this would consist of finding a group of domain experts and interview them to try to represent their most relevant knowledge, which itself was acquired from textbooks and other experts in a system of rules and symbols.[15] Modern approaches to knowledge acquisition include automated analysis of natural language in e.g. user manuals or textbooks and storing the acquired knowledge in (general-purpose) ontologies. As such, the principle of transferring symbolic knowledge from humans to artificial systems has not changed, but the method of this transfer has largely been automated.

Modern ontologies play an important role in allowing robots to perform actions in real-world environments. One of the problems in creating general-purpose robots is that the tasks they should perform are often greatly underspecified.[16] For example, a cooking robot trying to follow a recipe may encounter the instruction to "add a cup of water to the pan." While this is trivial for any human to follow, it requires knowledge about where to find water, where to get a cup from, and not to add the cup itself to the pan but merely its contents. This commonsense knowledge that is so self-evident to humans is not usually available to robots. However, online accessible ontologies such as Cyc may help by structuring commonsense knowledge in a symbolic, structured format available to robots and other computer systems.[17]

9 Avoiding exhaustive search that would be computationally prohibitive for anything but very small state spaces.
10 Or any of the river-crossing puzzles such as the related *jealous husbands problem* or the identical *foxes and chickens problem.*
11 Ohlsson 2012.
12 Laird 2012.
13 E.g. "What is the form of the individual organisms (e.g. lancet-shaped for cocci, fusiform for rods, etc.)?" from Buchanan and Shortliffe 1984.
14 Yu et al 1979.
15 Russell and Norvig 2010.
16 de Kleijn et al 2014.
17 Lenat et al 1990.

4 Artificial neural networks

However powerful symbolic AI had shown to be, by the end of the 1960s it became clear that there were some forms of human intelligence that it could not even begin to replicate. Interestingly, these seemed to be skills that human would perform effortlessly, such as object recognition, walking, or having a conversation.[18] As psychologist Steven Pinker noted: "The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard."[19] The tasks that AI research had concentrated on thus far were tasks that humans found particularly difficult, such as logic, symbolic algebra, and playing chess, which were seen as good indicators of intelligence. It was thought that as those problems were solved, the "easy" problems like vision would be solvable as well, an optimism well-illustrated by Herbert Simon's 1960 prediction that "machines will be capable, within twenty years, of doing any work a man can do."[20] As we now know, this would prove to be more difficult than expected. In 1973, reporting to the British Science Research Council, mathematician James Lighthill criticized the progress made in the field of AI in what is now known as the *Lighthill report.*[21] Specifically, the issue of combinatorial explosion in real-world problem solving was mentioned. Many of the problems that AI systems of the time were solvable for small toy problems, but turned out to be computationally intractable when scaled to real-world problems.[22] Also, it was unclear how problems such as vision, motor control, and noisy measurements such as encountered in robotics would be represented symbolically.

Some of these more difficult problems seemed to be particularly well-suited for another type of AI architecture known as *connectionism,* or *artificial neural networks (ANNs).* In these systems, information is not represented symbolically, but subsymbolically as activation values distributed over a network of elementary units with no intrinsic meaning.[23] Similar to neurons, which receive activity in the form of electrochemical signals through their dendrites and send activity along their axons, these units receive activation from other units, and send activation as a function of their input activation. Such networks are parameterized by the weights of the connections between the units and their activation functions, comparable to synapse strength and activation thresholds in the human brain. Although research into the mathematical modeling of neurons and their computational capability dates back to the 1940s when neuroscientists Warren McCulloch and Walter Pitts studied the implementation of logical functions in artificial neurons, it took almost half a century for artificial neural networks to take off.[24] In the 1980s, David Rumelhart and James McClelland published their now-standard collection *Parallel Distributed Processing,* in which they showed that artificial neural network models could account for a range of psychological phenomena, suggesting that the computational techniques they use are similar in nature. Such a network of artificial neurons can be used as a classifier, where it can take an input (such as an image), process it, and return an output such as a category label (is it a dog or a cat?). However, in order for it to do so, it needs to be trained. Most often, training such a neural network to classify dogs and cats is done using *supervised learning,* in which a large, correctly labeled dataset is presented to the network with a learning algorithm adjusting the network weights until it is able to correctly classify novel inputs.

At the start of the 2010s, deep neural networks started to reach or even surpass human performance

18 The observation that computers perform tasks that humans find difficult such as reasoning or playing chess very quickly and accurately, but have great difficulty performing tasks that humans find trivial such as walking or face recognition, is known as *Moravec's paradox.* Moravec (1990) argues that the difficulty for a computer system to solve a problem is proportional to the time evolution has had to optimize solving it in humans.

19 Pinker 1994, p 192.

20 Simon 1960, p 38.

21 Lighthill 1973.

22 These problems turned out to be most likely solvable only in exponential time where the required time to solve grows exponentially with input size, which is only acceptable for very small input sizes such as the toy problems AI was concerned with.

23 Although it could be argued that localist representations of the input and output layers of some connectionist models are symbolic in nature.

24 This is not to say that no important discoveries were made during the period in between, as important research into the power and limitations of neural networks was done by e.g. Minsky and Papert at MIT, and Rosenblatt at Cornell.

in image classification tasks, the start of a *deep learning* revolution.[25] Using a deep convolutional neural network and a training set of almost 40,000 images of 43 different German road signs,[26] researchers demonstrated a recognition accuracy of 99.5%, where humans scored 98.8%. At the same time, the same researchers showed 99.8% accuracy on the benchmark MNIST handwritten digit recognition dataset, which is near-human performance.[27] More recently, an ensemble of three deep neural networks trained to predict breast cancer from mammograms exceeded the average performance of six board certified radiologists. In another study, the researchers showed that in a so-called *double-reading process,* a process used in the UK for screening mammograms using two independent interpretations, the second reader's workload can be reduced by 88% without compromising the standard of care.[28] With these artificial neural networks reaching or even surpassing human-like performance, looking at the similarities between this type of artificial intelligence and its biological counterpart becomes even more interesting.

Human brains are not just 10^{11} neurons randomly crammed together in a skull, they are structured into a specific topology with some groups of neurons more densely connected than others. As mentioned earlier, the biological inspiration for artificial neural networks arose from the concept of a network of elementary units (neurons in humans), connected through weighted links (synapses in humans). These units are arranged in layers: an input layer representing the input to the network, an output layer representing the output, and one or more hidden layers. But determining how many layers we actually need to solve a certain problem is an art as well as a science. From a science perspective, some fundamental limitations have become clear. It was shown in 1969 that at least one hidden layer is necessary to learn some complex patterns, and sufficient to learn arbitrarily complex patterns.[29] However, the deep learning revolution that accompanied an impressive increase in the performance of artificial neural networks showed that adding more layers to a network can increase its performance. In these deep networks, higher layers represent more abstract features such as faces or letters, while lower layers represent more raw features such as edges or orientation. A similar architecture can be seen in the human visual cortex, where neurons in lower layers specifically respond to location and orientation while neurons in higher layers respond to faces or objects. In fact, the representations learned by deep networks show similarities to the representations developed in the primate visual system.[30] Such a hierarchical topology greatly increases representational power for a given number of parameters, both in biological and artificial neural networks.

When training ANNs using supervised learning, network weights are usually optimized using a technique known as *backpropagation.* The backpropagation algorithm computes the gradient of the error function[31] with respect to the network weights at the output layer, which is then propagated back to previous layers. This gradient information can then be used to adjust network weights, e.g. using gradient descent. While a powerful technique for supervised learning of network weights, as a model of brain function backpropagation was quickly criticized for being biologically implausible.[32] Biological neurons do not seem capable of transmitting information about the error gradient backwards along the axon, or any information at all for that matter. This is of course not to say that there are no return pathways in the brain, as there clearly are, but units or pathways that compare the output of a neuron to its required

25 The *deep* in *deep learning* refers to the number of layers (depth) in an artificial neural network, see below for an explanation.
26 The German Traffic Sign Recognition Benchmark (GTSRB) dataset containing more than 50,000 traffic sign images was used; Stallkamp et al 2011.
27 Cireşan et al 2012.
28 McKinney et al 2020.
29 More specifically, it was shown that learning non-linearly separable functions such as XOR requires at least one hidden layer, and that this is enough to approximate any continuous function; Minsky and Papert 1969, Cybenko 1989.
30 Yamins and DiCarlo 2016.
31 The error function defines the error between the actual output of the network and the required output of the network for a set of input–output pairs. For classification problems (a popular use of ANNs) cross entropy is often used.
32 And not only backpropagation, but the entire endeavor of connectionist modeling, see e.g. Crick 1989 for a scathing commentary.

output and propagate it across layers to cause changes in synaptic strength have not been discovered.[33] Not only is the biological plausibility of backpropagation questionable, the entire supervised learning process could be argued to be implausible. Humans are simply not provided with thousands of correctly labeled training data[34] for every object or concept they encounter, something that is required for the successful training of a deep network. Instead, humans seem to learn through trial-and-error, which is perhaps better modeled through a reinforcement learning paradigm. Modern deep reinforcement learning models use deep neural networks to approximate the expected outcome of possible actions to be taken, with impressive results.

In the animal visual cortex, neurons respond to the activation of other neurons in a specific area, known as the receptive field.[35] Convolutional neural networks are inspired by architecture of the animal visual cortex. Whereas in traditional artificial neural networks the units in each layer are connected to all the units in the previous layer,[36] in convolutional neural networks the units in a layer are connected to a *subset* of units in the previous layer. This greatly reduces the number of parameters of the network, reducing overfitting,[37] allowing for deeper networks, and reducing training time. Not only are convolutional neural networks successful classifiers, they can predict large-scale activation of human brain regions and the firing patterns of neurons, suggesting similar mechanisms and computational principles between the two.[38]

While deeper networks are more powerful, they are also harder to train and can suffer from the *vanishing gradient problem.*[39] Residual neural networks (ResNets) are inspired by the architecture of pyramidal cells in the cerebral cortex. In fully connected artificial neural networks, all units in a layer are connected to all units in the next layer. As such, there are no connections between units in layer x and layer $x+2$ or layer $x+3$. In a residual neural network, these connections are allowed, effectively skipping one or more layers when propagating activation (see Figure 1). It has been shown that for extremely deep networks, residual neural networks are easier to train, allow more layers, and perform better than non-residual networks given a specific network complexity.[40]

As said earlier in this section, artificial neural networks can now outperform human intelligence on certain specific tasks using techniques inspired by neurological principles. But even the types of networks that can outperform humans appear to have some idiosyncrasies that are remarkably different from human performance. So-called *adversarial examples* are inputs to a classifier that are slightly modified so that they are misclassified even though a human observer may see no difference.[41] It has recently been shown that the modification can be as small as a single pixel,[42] and does not have to be directly applied to the input data directly, but can also be applied to a real-world object that indirectly serves as an input, such as an object that is photographed.[43] These misclassifications can be quite stunning

33 It should be noted that the biological plausibility of backpropagation is controversial, and by no means a solved question. For example, there is evidence to suggest that when an action potential travels through an axon, it can cause a retrograde signal being sent to the presynaptic neuron through the dendrites. However, this is still far from actually propagating an error signal back across several neurons. See e.g. Stuart et al 1997, Bogacz et al 2000.

34 Training data in supervised learning consists of an input (e.g. a picture of a cat) and a desired output (e.g. the label "cat").

35 Hubel and Wiesel 1959.

36 This is referred to as a *fully connected* network.

37 *Overfitting* refers to the phenomenon where a network is trained to the point in which it can correctly classify the training data it has seen, without being able to generalize to novel instances. For example, it would be able to correctly classify its 10,000 training images as a cat, but fails to correctly classify a new picture of a cat.

38 Zhou and Firestone 2019.

39 The vanishing gradient problem occurs when the gradient of the error function becomes so small that network weights are no longer being updated. This is more likely to happen with very deep networks as the gradient decreases exponentially with the number of layers.

40 He et al 2016.

41 Although it should be noted that there are adversarial examples that fool both time-limited humans and computers, see e.g. Elsayed 2018; Goodfellow et al 2015.

42 Su et al 2019.

43 Kurakin et al 2017.

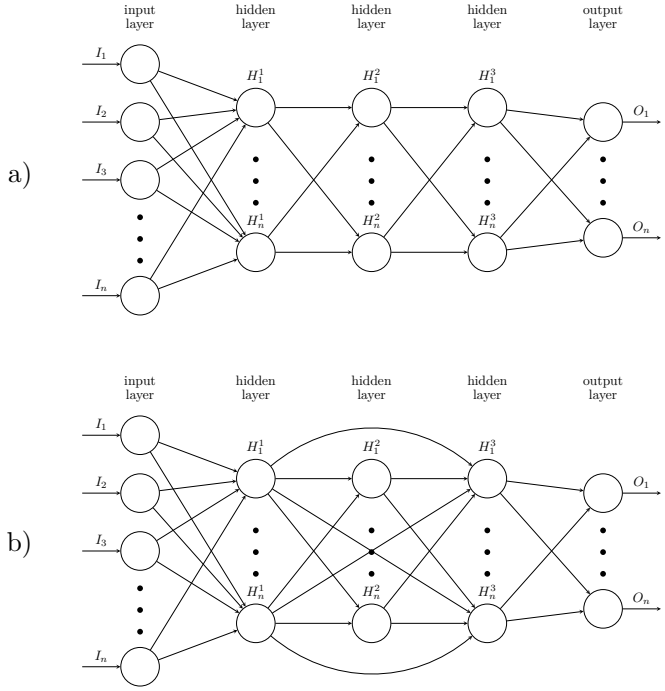

Figure 1. A regular deep neural network (a) compared to a residual neural network (b). Note that in (b) there are connections between units in layer H^1 and H^3, effectively skipping layer H^2.

to a human observer, for instance when a clear image of an elephant is being classified as a baseball, or a car as a milk can. Although interesting from a machine learning perspective, these findings are perhaps even more interesting from the perspective of human intelligence. Adversarial examples are often indistinguishable from their originals to humans, but are able to fool deep networks causing them to misclassify them. And not only that, deep networks assign high confidence ratings to their incorrect classification. This phenomenon casts doubt on the alleged similarity between deep neural networks and human object recognition mechanisms. However, some authors[44] have argued that these differences may not be caused by a qualitative difference between artificial and biological object recognition mechanisms and computational principles, but by the limitations of the human visual system which cannot perceive the perturbations used in adversarial examples. In other words, the existence of adversarial examples may not tell us anything about the high-level mechanisms of object recognition, but the low-level architecture of the visual system.[45] Perhaps the difference between human and computer intelligence in object recognition can be best illustrated with an analogy.[46] Human cognition allows us to distinguish between objects appearing to be like something and objects being something, for instance when distinguishing between a cloud that looks like a dog and an actual dog. Deep networks have no such luxury, and instead are forced to pick the label that is most likely.

5 Conclusion

The fields of artificial intelligence and the study of human intelligence have been intertwined, and devoting only one book chapter can hardly be enough. We limited ourselves here to two central forms of knowledge representation, symbolic and subsymbolic. The first approach represents knowledge symbolically, and

44 For example Zhou and Firestone 2019.
45 Although any evidence that the low-level architecture of the visual system is different for humans and computers should not come as a surprise; see Zhou and Firestone 2019.
46 Analogy taken from Zhou and Firestone 2019.

reasons using these symbols. Problem solving can be seen as symbol manipulation, according to this view. And knowing what we know about human cognition, it seems that at least part of our reasoning is symbolic in nature, not to speak of human symbolic communication.[47] On the other hand, it is also clear that many of the tasks we perform, such as vision and walking do not lend themselves well to symbolic representation. Artificial neural networks have found inspiration in the biological brain in many forms, not only the function of individual neurons, but also topological constraints such as deep, convolutional, and residual neural networks.

The idea of artificial neural networks is appealing. We know that the human brain produces some very intelligent behavior, so trying to emulate its mechanisms seems like an appropriate course of action. But opinions differ on whether brain-inspired artificial intelligence holds the key to creating truly intelligent artificial systems. It could be that although neuroscience has inspired ANN research, we have already reached the limits of what can be learned from brain research.[48] Although there are many commonalities between human intelligence and its artificial implementations, the one dimension on which they differ greatly is domain-specificity. While expert systems and deep networks can show better-than-human performance on several tasks, these remain very specific and are limited to the tasks these systems were designed or trained for. Although progress has been made in transfer learning and other areas, generalizability remains a puzzle and these developments have not yet been scaled to true out-of-domain performance. A computer system implementing artificial *general* intelligence[49] remains elusive, and although it has been the topic of research for decades, no big leaps in progress have been reported. The question remains whether human-level artificial intelligence—if ever achieved—will be the result of incremental progress on deep supervised, unsupervised and reinforcement learning, or that a paradigm shift is needed for artificial general intelligence. Meanwhile, the mechanisms causing human intelligence are not any less elusive. It seems that the one thing that is absolutely clear is that both the fields of artificial intelligence and cognitive science have a lot of work ahead of them. With the already impressive success of biologically inspired techniques, it is inevitable that new discoveries about the human brain and mind will further advance the state of artificial intelligence.

Bibliography

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. G. (2000). Frequency-based error back-propagation in a cortical network. IJCNN (5), 211–216. https://doi.org/10.1109/ijcnn.2000.857899

Chomsky, N. (1959). A review of B. F. Skinner's *Verbal Behavior*. Language, 35, 26–58.

Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition, abs/1202.2745, 3642–3649. https://doi.org/10.1109/cvpr.2012.6248110

Crick, F. (1989). The recent excitement about neural networks. Nature, 337, 129–132. https://doi.org/10.1038/337129a0

de Kleijn, R., Kachergis, G., & Hommel, B. (2014). Everyday robotic action: Lessons from human action control. Frontiers in Neurorobotics, 8, 13. https://doi.org/10.3389/fnbot.2014.00013

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018).

47 We have purposefully refrained from discussing human language (well, with the exception of one Chomsky reference) in this book chapter due to space constraints. Discussing it in the context of artificial intelligence would open up a can of worms that no single book, let alone a chapter, could do justice to, specifically in the light of the recent successes of large language models such as OpenAI's ChatGPT.

48 Zador 2019.

49 Artificial general intelligence (AGI) refers to a hypothetical type of AI that would be able to learn any task that humans can perform.

Adversarial examples that fool both computer vision and time-limited humans. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, 3914–3924.

Newell, A., Shaw, J. C., & Simon, H. A. (1957) Empirical explorations with the logic theory machine. Proceedings of the Western Joint Computer Conference, 15, 218-239.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. International Conference on Learning Representations, abs/1412.6572. http://arxiv.org/abs/1412.6572

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/cvpr.2016.90

Hershauer, J. C., & Simon, H. A. (1978). The new science of management decision. The Academy of Management Review, 3, 161. https://doi.org/10.2307/257591

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology, 148, 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. Artificial Intelligence Safety and Security, abs/1607.02533, 99–112. https://doi.org/10.1201/9781351251389-8

Laird, J. E. (2012). The Soar cognitive architecture. https://doi.org/10.7551/mitpress/7688.001.0001

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: Toward programs with common sense. Communications of the ACM, 33, 30–49. https://doi.org/10.1145/79173.79176

Lewicki, G., & Marino, G. (2003). Approximation by superpositions of a sigmoidal function. Zeitschrift Für Analysis Und Ihre Anwendungen, 2, 463–470. https://doi.org/10.4171/zaa/1156

McClelland, J. L. (2009). Is a machine realization of truly human-like intelligence achievable? Cognitive Computation, 1, 17–21. https://doi.org/10.1007/s12559-009-9015-x

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577, 89–94. https://doi.org/10.1038/s41586-019-1799-6

Minsky, M., & Papert, S. (1987). Perceptrons: An introduction to computational geometry. I–XV, 1–292.

Newell, A., & Simon, H. A. (1988). GPS, a program that simulates human thought. Readings in Cognitive Science, 453–460. https://doi.org/10.1016/b978-1-4832-1446-7.50040-6

Ohlsson, S. (2012). The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. The Journal of Problem Solving, 5. https://doi.org/10.7771/1932-6246.1144

Pilkington, A. (1996). The language instinct: The new science of language and mind. Language and Literature: International Journal of Stylistics, 5, 71–74. https://doi.org/10.1177/096394709600500107

Russell, S. J., & Norvig, P. (1995). Artificial intelligence: A modern approach. Choice Reviews Online, 33, 33-1577-33–1577. https://doi.org/10.5860/choice.33-1577

Sowa, J. (2000). Knowledge representation: Logical, philosophical, and computational foundations. Computational Linguistics, 27, 286–294. https://doi.org/10.1162/089120101750300544

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011). The German traffic sign recognition benchmark: A multi-class classification competition. The 2011 International Joint Conference on Neural Networks,

1453–1460. https://doi.org/10.1109/ijcnn.2011.6033395

Stuart, G., Spruston, N., Sakmann, B., & Häusser, M. (1997). Action potential initiation and backpropagation in neurons of the mammalian CNS. Trends in Neurosciences, 20, 125–131. https://doi.org/10.1016/s0166-2236(96)10075-8

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23, 828–841. https://doi.org/10.1109/tevc.2019.2890858

Sutherland, S., & Needham, R. (n.d.). Artificial intelligence: A general survey.

Swartout, W. R. (1985). Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project. Artificial Intelligence, 26, 364–366. https://doi.org/10.1016/0004-3702(85)90067-0

Truck, F., & Moravec, H. (1991). Mind children: The future of robot and human intelligence. Leonardo, 24, 242. https://doi.org/10.2307/1575314

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience, 19, 356–365. https://doi.org/10.1038/nn.4244

Yu, V. L. (1979). Antimicrobial selection by a computer: A blinded evaluation by infectious diseases experts. JAMA: The Journal of the American Medical Association, 242, 1279–1282. https://doi.org/10.1001/jama.242.12.1279

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. Nature Communications, 10. https://doi.org/10.1038/s41467-019-11786-6

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. Nature Communications, 10. https://doi.org/10.1038/s41467-019-08931-6