The Foundations of Deep Learning with a Path Towards General Intelligence

Abstract

Like any formal field of science, AI may be approached axiomatically. We formulate requirements for a general-purpose, human-level AI system in terms of postulates. We review the methodology of deep learning, examining the explicit and tacit assumptions in deep learning research. Deep Learning methodology seeks to overcome limitations in traditional machine learning research as it combines facets of model richness, generality, and practical applicability. The methodology so far has produced outstanding results due to a productive synergy of function approximation, under plausible assumptions of irreducibility and the efficiency of back-propagation family of algorithms. We examine these winning traits of deep learning, and also observe the various known failure modes of deep learning. We conclude by giving recommendations on how to extend deep learning methodology to cover the postulates of general-purpose AI including modularity, and cognitive architecture. We also relate deep learning to advances in theoretical neuroscience research.

1 Introduction

Deep learning is a rapidly developing branch of machine learning which is clustered around training deep neural models with many layers and rich computational structure well suited to the problem domain [Goodfellow *et al.*, 2016; Schmidhuber, 2015]. Initially motivated by modelling the visual cortex [Fukushima, 1980; Fukushima, 2013], human-level perceptual performance was approached and eventually attained in a number of challenging visual perception tasks such as image recognition with the aid of GPU acceleration [LeCun *et al.*, 1989; Ranzato *et al.*, 2007; Graves *et al.*, 2009; Ciresan *et al.*, 2011]. The applications quickly extended to other computer vision tasks such as image segmentation [Ciresan *et al.*, 2012], producing a variety of impressive results in visual information processing such as style transfer [Gatys *et al.*, 2016], opening new vistas in machine learning capabilities. The applications have been extended to domains beyond vision, such as speech recognition [Graves *et al.*, 2013], language processing [Kim *et al.*, 2016], and reinforcement learning [Koutník *et al.*, 2013; Mnih and others, 2015], often with striking performance, proving the versatility and the significance of the approach in AI, urging us to consider whether the approach may yield a general AI (called Artificial General Intelligence (AGI) in some circles), and if so which problems would have to be tackled to make deep learning approach truly human-level AI that covers all aspects of cognition.

We analyze the approach from a 10.000 feet vantage point, revisiting the idea of AI axiomatization. Although, we are generally in agreement with Minsky that the attempt to make AI like physics is likely a futile pursuit, we also note the achievements of later theorists who have applied Bayesian methods successfully. We make no attempt to formalize any of our claims due to space consideration, however we discuss relevant research in cognitive sciences. Then, we apply the same foundational thinking to deep learning critically probing its intellectual foundations. The axioms, or postulates, of AI, are examined with an eye towards whether the current progress in deep learning in some way satisfies them, and what has to be done to fill the gap. The present paper may thus be regarded as an analytical, critical meta-level review, rather than a comprehensive review such as [Schmidhuber, 2015].

2 Postulates of General AI

One of the most ambitious mathematical models in AGI research is AIXI [Hutter, 2007] which is a universal Reinforcement Learning (RL) model that can be applied to a very large variety of AI agent models and AI tasks including game playing, machine learning tasks, and general problem solving. AIXI is based on an extension of Solomonoff's sequence induction model [Solomonoff, 1957; Solomonoff, 2008] which works with arbitrary loss and alphabet [Hutter, 2003], making the aforementioned induction problem fairly general. Hutter proves in his book [Hutter, 2005] that many problems can be easily transformed to this particular formulation of universal induction. There are a few conditions that have to be

satisfied for a system to be called a universal induction system, and even then the system must be realized in a practical manner so as to be widely applicable and reproduce the cognitive competencies of homo sapiens, or failing that, a less intelligent animal.

The AIXI model combines Bellman equation with universal induction, casting action selection as the problem of maximizing expected cumulative reward in any computable environment. Although RL is a common approach in machine learning, AIXI had the novelty that it focused solely on universal RL agents. When viewed this way, it is obvious that AIXI is a minimalist cognitive architecture model, that exploits the predictive power of induction in RL setting, that does give the model the kind of versatility noted above. Solomonoff induction presents a desirable limit of inductive inference systems, since it has the least generalization error possible; the error is dependent only on the stochastic source and a good approximation can learn from very few examples [Solomonoff, 1978]. AIXI model also retains a property of optimal behavior, Hutter deliberates that the model defines optimal, but incomputable intelligence, and thus any RL agent must approximate it. Therefore, our axiomatization must consider the conditions for Solomonoff's universal induction model, and consequently AIXI, to be approximated well, but we believe additional conditions are necessary for it to also satisfy generality in practice and within a versatile system, as follows:

Completenes The class of models that can be acquired by the machine learning system must be Turing-complete. If a large portion of the space of programs is unavailable to the system, it will not have the full power and generalization properties of Solomonoff induction. The convergence theorem in that case is voided, and the generalization performance of Solomonoff induction cannot be guaranteed [Solomonoff, 1978].

Stochastic Models The system requires an adequately wide class of stochastic models to deal with uncertainty in the real world, a system with only deterministic components will be brittle. Induction is better suited to working with stochastic models, one example of such an approach is Wallace's Minimum Message Length (MML) model where we minimize the message length that contains both the length of the statistical model encoding and data encoding length relative to model [Wallace and Boulton, 1968; Wallace and Dowe, 1999].

Bayesian Prediction The system must compute the inferences with Bayes' law. The inference in Solomonoff's model is considered Bayesian. In neuroscience, the Bayesian Brain Hypothesis has been mostly accepted, and the brain is often regarded as a Bayesian inference machine that extracts information from the environment in theoretical neuroscience. Jaynes introduced the possibility of Bayesian reasoning in the brain from a statistical point of view [Jaynes, 1988]. The Bayesian approach to theoretical neuroscience is examined in a relatively recent book [Doya *et al.*, 2007]. Fahlman et. al introduced the statistically motivated energy minimizing Boltzmann machine model [Fahlman *et al.*, 1983]; Hinton et. al connected the induction principle of Minimum Description Length and Helmholtz free energy introducing the autoencoder model in 1993 [Hinton and Zemel, 1993]. Bialek's lab has greatly contributed to the understanding of the Bayesian nature of the brain, a decent summary of the approach detailing the application of the information bottleneck method may be found in [Bialek *et al.*, 2001]. Friston has later rigorously applied the free energy principle and has obtained even more encouraging results, he explains the Bayesian paradigm in [Friston, 2012]. Note that Helmoltz free energy and the free energy principle are related, and both are related to approximate bayesian inference.

Principle of Induction The system must have a sound principle of induction that is equivalent to Solomonoff's model of induction which uses an a priori probability model of programs that is inversely and exponentially proportional to program size. Without the proper principle of induction, generalization error will suffer greatly, as the system will be corrupted. Likewise, as Solomonoff induction is more completely approximated, the generalization error will decrease dramatically, allowing the system to obtain one-shot learning first predicted by Solomonoff, achieving a successful generalization from a sufficiently complex single example without any prior training whenever such an example is possible.

Practical Approximation Solomonoff induction has an exponential worst-case bound with respect to program size rendering it infeasible. This surely is not a practical result, any approximation must introduce algorithmic methods to obtain a feasible approximation of the theoretical inductive inference model.

Incremental Learning The system must be capable of cumulative learning, and therefore it must have a model of memory with adequate practical algorithms. Solomonoff has himself described a rather elaborate approach to transfer learning [Solomonoff, 1989], however, it was not until much later that experimental results were possible for universal induction since Solomonoff's theoretical description did not specify an efficient algorithm. The first such result was obtained in OOPS system [Schmidhuber, 2004] demonstrating significant speedups for a universal problem solver.

Modularity and Scalability The system must be composed of parametrized modules that attend to different tasks, allowing complex ensemble systems to be built for scalability like the neocortex in the human brain. A monolithic system is not likely to scale well, the system must be able to adapt modules to distinct tasks, and then be able to re-use the skills. A modular system also provides a good base for specialization according to modality and cognitive task, starting from a common module description. In the human brain, there are both functional regions and a complex, hierarchical modular structure in the form of cortical columns, and microcolumns.

Cognitive Architecture The system must have a cognitive architecture, depending on modularity that will address typical cognitive functions of learning, memory, perception, reasoning, planning, and language as well as aspects of robotics which allow it to control robotic appendages. This manner of organization is modeled after the human brain, however, it seems essential for any real-world AI system that requires these basic competencies to deliver robust performance across a sufficiently general set of cognitive tasks. Even if unlike the brain, the system must have an architectural design, or one that is capable of introducing the required architecture.

These reasonable and desirable properties of a complete AI system lead naturally to a top-down design sometimes called an AGI Unification Architecture among practitioners, if built around the floor plan of a universal induction system such as AIXI. An example of such an approach to designing a cognitive architecture may be seen in [Potapov *et al.*, 2016]. However, this is not necessarily the only kind of solution. An adequate architecture could also be built around a deep learning approach; let us therefore proceed to its postulates.

3 Postulates of Deep Learning

Deep Learning is a particular kind of Artificial Neural Network (ANN) research which shares some commonalities and inherits some assumptions / principles from earlier ANN research some of which may seem implicit to outsiders. We try to recover these tacit or implicit assumptions for the sake of general AI readership, and also delineate the borders of deep learning from other ANN research in the following:

No Free Lunch The well-known No Free Lunch theorem for machine learning [Wolpert, 1996] implies that there can be no general learning algorithm that will be effective for all problems. This theorem has generated a tendency towards model-based learning in ANN research where the researcher tries to design a rich network model that covers all contingencies in the domain but uses insights into the problem domain and thus the experiment does not suffer from the unreasonable large search space of a model-free learning method. From image processing to language, this particular blend of specificity and generality seems to have guided deep learning quite successfully and resulted in impressive outcomes. The specificity determined by the ANN researcher may be likened to innateness in cognitive science. Note that AGI theorists have argued otherwise [Everitt *et al.*, 2014], therefore this heuristic principle remains arguable.

Epistemic Non-reductionism This is the view that loosely depends on Quine's observation that epistemic reductionism often fails in terms of explanatory power for the real world [Quine, 1951]. When we look at a deep learning vision architecture, we see that the irreducible patterns of visual information are indeed stored as they are useful however not overmuch; the system does not store every pattern much like our brains. Epistemic irreducibility is a guiding principle in deep learning research, and it is why deep learning models are large rather than small and minimalistic as in some ANN research.

Eliminative Materialism Churchland's philosophical observation that the brain does not deal in any of the folk psychological concepts in cognitive science literature, but must be understood as the activation state and trajectory of the brain [Churchland, 1981], plays a fundamental intellectual role in the deep learning approach, where we shift our attention to brain-like representations and learning for dealing with any problem, even if it looks like a matter of propositional logic to us.

Subsymbolic & Distributed Representation Expressed in detail in the classical connectionist volume [Rumelhart *et al.*, 1986], this principle is the view that all representations in the brain have a distributed, real-valued representation rather than discrete, symbolic representations that computer scientists prefer in their programs. Sparse Coding hypothesis has been mostly confirmed in neuroscience, therefore we do know that the brain uses population codes that are sparse, distributed, and redundant. Unlike a symbolic representation, the brain networks are fault-tolerant and redundant, and deal with uncertainty at every level. Subsymbolic representations are more robust and better suited to the nature of sensory input. However, we also know that grandmother cells exist which may correspond to predicates, which are still best modeled as non-linear detectors, or ReLu units, in a neural network.

Universal Approximation The universal approximation theorem [Hornik, 1991] for multi-layer feed forward neural networks underlies the heuristic of using many hidden layers in a deep learning architecture. The theorem shows that a multi-layer neural network can approximate arbitrary continuous real-valued functions. Therefore, the system is capable of representing any mapping under mild assumptions, including those with irregular features forming a synergy with the epistemic non-reductionism postulate.

Deep Models The number of layers in a feed forward network, or the circumference of a Recurrent Neural Network (RNN) must be greater than 3, meaning multiple hidden layers in a multi-layer feed forward network, or an RNN with complex topology. Model depth avoids much of the criticism in Minsky and Papert's critical book on neural networks that showed perceptrons cannot learn concave discriminants [Minsky and Papert, 1969], and its later editions that extend the criticism to multi-layer models. In today's ANN applications we observe all manners of intricate discrimination models were successfully learnt, however shallow networks will still not avoid Minsky's observations. A complexity analysis also supports that increasing depth can result in asymptotically smaller networks for the same function representation [Telgarsky, 2016], implying that deep models are fundamentally more efficient.

Hierarchy and Locality A distinguishing feature of deep learning is that it contains local pattern recognition networks and a hierarchy of these pattern recog-

nition circuits that affixes the local and global views. Thus, a sequence of convolutional and pooling layers have been a staple of image processing applications in deep learning as the convolutional layer is basically a set of texture recognition patches, and downsampling via max-pooling gives us a dimensionality reduction and the ability to hierarchically combine pattern recognizers efficiently. This organization was inspired by 2d image processing in the visual cortex, however many domains can benefit from the same organizational principle since they apply to any sensory array. The principle is also valid for domains that are not directly sensory arrays, but maintain a similar topological relation. The principle also has great synergy with the depth principle because the network tries to capture perceptually salient features and avoids learning irrelevant patterns making it possible to increase network depth which avoids Minskyan objections even more effectively.

Gradient Descent Perhaps the most common feature of deep learning is that a variation of back propagation or gradient descent is used to train the model. This is required since any other way to train the large networks in deep learning research would be infeasible. Other methods such as variational learning and MCMC tree search have been applied in deep learning research, however this principle has remained fairly constant as it is necessitated by other principles above, which may result in billions of real valued parameters to be trained.

Dataflow models & SIMD acceleration Since the number of parameters to be trained is large, exploiting data-parallelism through SIMD-based accelerators such as GPU's, and later executing data-flow representations on FPGA's have proven to be an essential factor for deep learning research. This property of deep learning corresponds to the massive parallelism property of the brain.

4 Shortcomings of Deep Learning

Although deep learning has generated phenomenal results, it also has some shortcomings that are being worked on. The most common limitation is that a typical deep learning architecture requires on the order of 10,000 or more examples. Some of the largest experiments have used millions of examples, therefore this was simply not an issue that was focused on. It may well be the case that this is a fundamental shortcoming of deep learning, however, researchers have tried solutions such as using stochastic gradient over the entire set of samples, as a usual statistical approach would necessitate, instead of running BP in epochs, which imitates the brain's online learning capability. Another common problem is that most deep learning uses supervised learning, which presents a problem in terms of constructing many labeled/annotated examples for every new problem. Autoencoder [Hinton and Zemel, 1994] is an unsupervised learning model, and it has many variations and applications in deep learning, however, most applications still require a good deal of hand crafted data. A strange problem persists in deep learning systems, which makes them easy to fool in ways that are not intuitive to humans, such as a simple perturbation causing a misclassification, an intuitively unrelated artificial image recognized as a natural image, or a specially crafted patch on an unrelated image causing a misclassification. These might either be symptoms of fundamental limitations, or they might be ameliorated with better deep learning models. We observe that these issues look much like overfitting, i.e., poor generalization performance.

5 Extending Deep Learning

When we contrast the general AI postulates and deep learning postulates, we see some interesting overlap and also some areas where deep learning requires a good deal of development. A deep learning system has one sort of completeness that stems from the universal approximation theorem, and dataflow models can be augmented with arbitrary computational units such as the Neural Turing Machine model [Graves *et al.*, 2014], and the later Differentiable Neural Computer model [Graves and others, 2016] that augments neural networks with external memory. Program class extensions of this sort may be an integral part of next-generation deep learning. Recent proposals for non-Euclidian embedding of data also enhance generality of deep learning models [Bronstein *et al.*, 2017].

It is possible to design deep architectures for rigorous stochastic models, which is an important extension to deep learning that will increase robustness.

Typically, deep learning lacks a principle of induction, but at the same time a stochastic model of induction is implicit in deep learning as the information bottleneck analysis of deep learning shows [Tishby and Zaslavsky, 2015], where we can view deep learning as a lossy compression scheme that forgets unnecessary information. Such theories will lead to better generalization performance. [Kawaguchi *et al.*, 2017] applies random matrix theory to generalization in deep learning, and introduces a new regularization method for improving generalization.

Progressive deep learning architectures add layers as necessary, substantiating an important analogy to SVM's function class iteration [Rusu *et al.*, 2016]. Much richer forms of induction may be beneficial for improving a deep learning network's generalization power. The training procedure in deep learning is efficient but only locally optimal, in the future a combination of neuro-evolution and gradient descent may outperform gradient descent and approximate universal induction better. Evolution has already been applied to automated design of deep networks [Miikkulainen *et al.*, 2017; Petroski Such *et al.*, 2017]. Neuro-evolution has been shown to be effective in game playing [Risi and Togelius, 2017] and other tasks that are difficult for deep learning, and therefore it might displace deep learning methodology altogether in the future.

Deep learning architectures gained memory capability with the LSTM unit, and similarly designed memory

cells, however, long-term memory across tasks remains problematic. A good realization of algorithmic memory in deep learning is Neural Task Programming (NTP) [Xu *et al.*, 2017] which achieves an indexical algorithmic memory based on LSTM and the ability to hierarchically decompose skills which has been successfully applied to robotics tasks. Progress in the direction of NTP is likely to be a major improvement for deep learning, since without cumulative and hierarchical learning intelligence is highly restricted.

Recently, progress has been made in the matter of modularity with Hinton's update of Capsule Networks, that models the cortical architecture for visual tasks [Sabour *et al.*, 2017]. Capsule Networks adds dynamic routing between visual processing modules with affine transformations, enhancing invariance and defines neural modules as capsules that may be arranged like neurons. Capsules correspond to visual entities in the model, therefore capsules that recognize a face decompose into eyes, a nose, lips, and so forth. The step from monolithic to modular deep learning is as powerful as the step from shallow to deep networks, hence this line of research is a significant extension of deep learning. A similar line of research is advanced by Vicarious, which propose a recursive neural architecture that exploits lateral connections accounting for distinct feature sets such as contour and surface, and the hierarchical representation of entities like in Capsule Networks [George *et al.*, 2017]; their system can reportedly break CAPTCHA's. Hawkins proposes a new cortex architecture that introduces pyramidal neurons, active dendrites, and multiple integration sites, identifying cortical computations for hierarchical sequence memory, and it intriguingly involves dendritic computation [Hawkins and Ahmad, 2016]. Capsule Networks might be enhanced to provide a similar dendritic model eventually, or capsule-like speciation might be ported to Hawkins's model.

Cognitive architectures built on symbolic concepts may not be readily applicable to deep learning, however, modeling the functional anatomy of the brain creates much needed synergy with neural networks. For instance, in Deep Mind's I2A model [Weber and others, 2017], we see a direction towards capturing more brain function in the form of imagining future states, while PathNet presents a modular, reflective learning system that can recombine network modules by evolving paths over the network [Fernando *et al.*, 2017]. Both neural architectures exhibit progress towards a more complete cognitive neural architecture. Another recent direction is the relational networks that model reasoning [Santoro *et al.*, 2017]. Conceivably, neural models of fundamental cognitive functions may be developed with a similar methodology, and bound in a connectionist agent architecture. Likewise, the active inference agent of [Friston *et al.*, 2017] with deep temporal models captures the essentials of functional anatomy based on hierarchical probabilistic models, and even gives us a fully unsupervised agent model that is quite intriguing from a scientific perspective.

6 Discussion and Future Research

Despite recent criticism raised against deep learning [Marcus, 2018], almost all of the postulates of general AI we have outlined seem achievable, however, with major improvements over existing systems. While it is entirely possible for a traditional symbolic-oriented system to achieve the same performance, the advantages of deep learning approach cannot be neglected, and the possible extensions to deep learning discussed may also ameliorate the common shortcomings we summarized. Another combination that might work is the combination of the symbolic AI approach with deep learning. In some circles, researchers pursue a mathematical AI unification approach (like AIXI approximations), however, the merits of such an approach are yet to be proven experimentally over deep learning. It seems prudent to at least try to integrate deep learning faithfully in existing AI architectures, or for new architectures, attempt to construct them solely on a neural architecture. In the future, we expect a convergence of more powerful training methods and deep architectures, taking us to a more model-free learning system, and more capable, modular neural agent architectures inspired by neuroscience.

References

[Bialek *et al.*, 2001] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[Bronstein *et al.*, 2017] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34:18–42, July 2017.

[Churchland, 1981] Paul M. Churchland. Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78(February):67–90, 1981.

[Ciresan *et al.*, 2011] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. A committee of neural networks for traffic sign classification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1918–1921, 2011.

[Ciresan *et al.*, 2012] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Neural networks for segmenting neuronal structures in EM stacks. In *ISBI Segmentation Challenge Competition: Abstracts*, 2012.

[Doya *et al.*, 2007] K. Doya, S. Ishii, A. Pouget, and R. P. N. Rao. *Bayesian brain: Probabilistic approaches to neural coding*. The MIT Press, 2007.

[Everitt *et al.*, 2014] Tom Everitt, Tor Lattimore, and Marcus Hutter. Free lunch for optimisation under the universal distribution. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2014, Beijing, China, July 6-11, 2014*, pages 167–174, 2014.

[Fahlman *et al.*, 1983] Scott E. Fahlman, Geoffrey E. Hinton, and Terrence J. Sejnowski. Massively parallel architectures for ai: Netl, thistle, and boltzmann machines. In *Proceedings of the Third AAAI Conference on Artificial Intelligence*, AAAI'83, pages 109–113. AAAI Press, 1983.

[Fernando *et al.*, 2017] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. PathNet: Evolution Channels Gradient Descent in Super Neural Networks. *ArXiv e-prints*, January 2017.

[Friston *et al.*, 2017] Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, 77:388 – 402, 2017.

[Friston, 2012] Karl Friston. The history of the future of the bayesian brain. *Neuroimage*, 62(248):1230–1233, 2012.

[Fukushima, 1980] K. Fukushima. Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[Fukushima, 2013] Kunihiko Fukushima. Artificial vision by multi-layered neural networks: Neocognitron and its advances. *Neural Networks*, 37:103–119, 2013.

[Gatys *et al.*, 2016] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2414–2423, 2016.

[George *et al.*, 2017] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, A. Lavin, and D. S. Phoenix. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 2017.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[Graves and others, 2016] Alex Graves et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[Graves *et al.*, 2009] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for improved unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5), 2009.

[Graves *et al.*, 2013] Alex Graves, Abdel-Rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649. IEEE, 2013.

[Graves *et al.*, 2014] A. Graves, G. Wayne, and I. Danihelka. Neural Turing Machines. *ArXiv e-prints*, October 2014.

[Hawkins and Ahmad, 2016] Jeff Hawkins and Subutai Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits*, 10:23, 2016.

[Hinton and Zemel, 1993] Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 3–10, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[Hinton and Zemel, 1994] G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length and Helmholtz free energy. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems (NIPS) 6*, pages 3–10. Morgan Kaufmann, 1994.

[Hornik, 1991] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 4(2):251–257, March 1991.

[Hutter, 2003] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003. (On J. Schmidhuber's SNF grant 20-61847).

[Hutter, 2005] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

[Hutter, 2007] Marcus Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel and C. Pennachin, editors, *Artificial General Intelligence*, Cognitive Technologies, pages 227–290. Springer, Berlin, 2007.

[Jaynes, 1988] E. T. Jaynes. How does the brain do plausible reasoning? In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, volume 1, 1988.

[Kawaguchi *et al.*, 2017] K. Kawaguchi, L. Pack Kaelbling, and Y. Bengio. Generalization in Deep Learning. *ArXiv e-prints*, October 2017.

[Kim *et al.*, 2016] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2741–2749, 2016.

[Koutník *et al.*, 2013] Jan Koutník, Giuseppe Cuccu, Juergen Schmidhuber, and Faustino Gomez. Evolving large-scale neural networks for vision-based TORCS. In *Foundations of Digital Games*, pages 206–212, Chania, Crete, GR, 2013.

[LeCun *et al.*, 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Back-propagation applied to handwritten zip

code recognition. *Neural Computation*, 1(4):541–551, 1989.

[Marcus, 2018] G. Marcus. Deep Learning: A Critical Appraisal. *ArXiv e-prints*, January 2018.

[Miikkulainen *et al.*, 2017] R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy, and B. Hodjat. Evolving Deep Neural Networks. *ArXiv e-prints*, March 2017.

[Minsky and Papert, 1969] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969.

[Mnih and others, 2015] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Petroski Such *et al.*, 2017] F. Petroski Such, V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, and J. Clune. Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning. *ArXiv e-prints*, December 2017.

[Potapov *et al.*, 2016] Alexey Potapov, Sergey Rodionov, and Vita Potapova. Real-time ga-based probabilistic programming in application to robot control. In *Artificial General Intelligence - 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*, pages 95–105, 2016.

[Quine, 1951] W.V.O. Quine. Two dogmas of empiricism. *The Philosophical Review*, 60:20–43, 1951.

[Ranzato *et al.*, 2007] M. A. Ranzato, F. Huang, Y. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR'07)*, pages 1–8. IEEE Press, 2007.

[Risi and Togelius, 2017] Sebastian Risi and Julian Togelius. Neuroevolution in games: State of the art and open challenges. *IEEE Trans. Comput. Intellig. and AI in Games*, 9(1):25–41, 2017.

[Rumelhart *et al.*, 1986] David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA, 1986.

[Rusu *et al.*, 2016] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[Sabour *et al.*, 2017] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *NIPS*, pages 3859–3869, 2017.

[Santoro *et al.*, 2017] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy P. Lillicrap. A simple neural network module for relational reasoning. *CoRR*, abs/1706.01427, 2017.

[Schmidhuber, 2004] Juergen Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54:211–256, 2004.

[Schmidhuber, 2015] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[Solomonoff, 1957] Ray J. Solomonoff. An inductive inference machine. In *IRE National Convention Record, Section on Information Theory, Part 2*, pages 56–62, New York, USA, March 1957.

[Solomonoff, 1978] Ray J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Trans. on Information Theory*, IT-24(4):422–432, July 1978.

[Solomonoff, 1989] Ray J. Solomonoff. A system for incremental learning based on algorithmic probability. In *Proceedings of the Sixth Israeli Conference on Artificial Intelligence*, pages 515–527, Tel Aviv, Israel, December 1989.

[Solomonoff, 2008] Ray J. Solomonoff. Three kinds of probabilistic induction: Universal distributions and convergence theorems. *The Computer Journal*, 51(5):566–570, 2008.

[Telgarsky, 2016] Matus Telgarsky. benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1517–1539, 2016.

[Tishby and Zaslavsky, 2015] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, pages 1–5, 2015.

[Wallace and Boulton, 1968] Chris S. Wallace and David M. Boulton. A information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

[Wallace and Dowe, 1999] C. S. Wallace and D. L. Dowe. Minimum message length and kolmogorov complexity. *The Computer Journal*, 42(4):270–283, 1999.

[Weber and others, 2017] T. Weber et al. Imagination-Augmented Agents for Deep Reinforcement Learning. *ArXiv e-prints*, July 2017.

[Wolpert, 1996] David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

[Xu *et al.*, 2017] D. Xu, S. Nair, Y. Zhu, J. Gao, A. Garg, L. Fei-Fei, and S. Savarese. Neural Task Programming: Learning to Generalize Across Hierarchical Tasks. *ArXiv e-prints*, October 2017.