Applying Deutsch's concept of good explanations to artificial intelligence and neuroscience - an initial exploration

Daniel C. Elton[1]

(1) Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA daniel.elton@nih.gov

Abstract. Artificial intelligence has made great strides since the deep learning resolution, but AI systems still struggle to discover generalizable principles and rules which allow them to extrapolate beyond their training data. For inspiration we look to a the domain of science, where the scientific method over many iterations has led to theories which show remarkable ability to extrapolate and sometimes even predict the existence new never-before-seen phenomena. According to David Deutsch, this type of extrapolation, which he calls "reach", is due to scientific theories being an example of a larger class of explanations he calls "good explanations", which are defined by the property that they are hard-to-vary. In this work we investigate Deutsch's hard-to-vary principle and how it relates to more formalized principles in deep learning such as the bias-variance trade-off and Occam's razor. Recent work suggests inductivist methods which involve fitting highly parametrized (easy-to-vary) models to big data can go quite far and may underlie much of how the brain learns. At the same time here we argue that Deutsch's principle may be a necessary additional component for achieving AI which is capable of generating good explanations of the world.

Keywords: Deep learning, artificial intelligence, generalization, extrapolation, Occam's razor, parsimony, critical rationalism, induction

1 Introduction

The field of Artificial Intelligence has made great strides since the ascendancy of deep learning since 2012. Deep learning models can now match or exceed human-level performance on natural image classification,[1] medical image diagnosis,[2] and the game of Go.[3] Several companies have fully autonomous vehicles on the road, and Google's Duplex system has wowed an audience with its ability to engage in natural language conversation. More recently the GPT3 model[4] has been shown to be able to write very convincing stories and give sensible sounding answers to questions. Yet there are still many things present day AI cannot do very well. Today's AI systems lack human-level common sense understanding, are clumsy at the robotic manipulation of objects, and lack the ability

to engage in casual reasoning. Another issues is that today's AI cannot learn from a few examples like humans and requires massive amounts of data to train. Most importantly though, today's AI systems are all narrow - they can only perform exactly the tasks they were trained to do, working within the distribution of their training data. As soon as today's AI systems are asked to work outside their training data distribution, they typically fail. Despite these shortcomings, we believe the deep learning paradigm (brute force fitting to large datasets) can go a long way as compute and data become ever more plentiful. However, as we will argue here, certain forms of extrapolation beyond the data fed into the system will always be out of reach to deep learning. The type of extrapolation ability we refer to is best demonstrated by scientific theories. According to David Deutsch, this type of extrapolation is enabled by scientific theories being an example of a larger class of explanations called "good explanations" which are defined by the property that they are hard-to-vary (hear-after denoted as HTV).[5] After introducing Deutsch's principle we explore how important it may be for both human and artificial intelligence. To the authors knowledge this research direction has not been undertaken previously in the published literature. However, as the author was finishing this article it was discovered Dennis Hackethal published a book in June 2020 exploring similar ideas.[6] It was also discovered there are a few people writing about Deutsch's ideas and AI online including Bruce Nielson and Elliot Temple. Their blogs are http://fourstrands.org/ and http://curi.us/, respectively. For a good general overview and critique of Deutsch's views on artificial intelligence and artificial creativity see Davenport (2016).[7]

2 Generalization vs extrapolation

In our view, the way the term "generalization" is used across both statistics and machine learning can often be misleading. In both fields,"generalization ability" is defined as the gap between test set error and training set error, where the training set and test set are drawn from the same distribution.[8,9] What we are interested in here is the ability to generalize *across* distributions, not within them, so for sake of clarity we will refer to this as "extrapolation". Deep learning systems are notoriously bad at extrapolation, often failing spectacularly when small distributional shifts occur. To give a few high profile examples, a deep learning system for diagnosing retinopathy developed by Google's Verily Life Sciences which reached human-level diagnostic ability in the lab performed very poorly in field trials in Thailand due to poor lighting conditions, lower resolution images, and images which had been stitched together.[10] A trained diagnostician would have been able to adapt to these conditions, but deep learning could not. Another high profile example occurred in 2015 when a deep learning system embedded in Google's Photos service was found to be tagging African American people as "gorillas". The quick-fix solution was to remove the category of gorillas. Three years later, *WIRED* magazine reported that the category was still missing, suggesting that the problem was not easy to solve.[11] In 2017 a much-lauded DeepMind deep reinforcement learning system[12] which achieved super-human

performance on several Atari games was shown to fail if minor changes are made, such as moving the paddle 3% higher in the game of Breakout.[13] This shows that the system has no grasp of basic physics and is built on top of a highly brittle fitting procedure. As a final example consider a deep reinforcement learning system for constructing molecules developed by a team from Insilico Medicine and Harvard which received much media fanfare after one of the molecules it generated was synthesized and shown to be active against cancer in a mouse model.[14] As pointed out in a critique by Walters & Murcko, the molecule it generated is very similar to two existing drug molecules which were in the system's training database.[15] In general deep generative models struggle to generate novel molecules and often start to generate highly nonphysical nonsense molecules as soon as they are asked to work outside their training set distribution (or "move off the data manifold").[16]

The range of input space within which a model, theory, or explanation can make accurate predictions has different names in different disciplines. The physicist David Deutsch calls it the "reach", philosophers sometimes call it the "explanatory power", and in some sub-fields of machine learning it is called the "applicability domain". As we discussed in detail by Hasson et al.[17] and by this author in a prior work,[18] the double descent phenomena indicates that machine learning models operate largely by local interpolation over a dense set of training data rather than by extracting global trends or extrapolatable rules. In light of these findings, the failures noted above are not so surprising.

What is missing from these systems? We believe we can gain insight into this question by turning our attention away from the types of statistical modeling done in mathematics and computer science departments to the type of modeling done in science departments, which is based theories developed using the scientific method. Physicists in particular have an impressive track record of being able to generate models that predict phenomena their creators had never seen or imagined before. For instance, Isaac Newton developed his laws of motion, which have been applied successfully to make predictions in many domains Newton had no direct experience in. In 1915 Einstein predicted that given the initial conditions of a light ray traveling close to the sun, the light ray would follow a curved trajectory, a finding that was confirmed experimentally a five years later during a solar eclipse.[19] Many radical and surprising phenomena have been predicted by physicists in advance, before any empirical confirmation - black holes, gravitational waves, antimatter, the Higgs boson, quantum computation, and metamaterials which can bend light around objects. Finally we note that much of today's high technology, such as novel drugs or spacecraft designs are first developed *in-silico* using physics-based modeling.

It is not our intention to wade into the philosophical debate about what denotes an "explanation" versus a "pure prediction". In what follows we treat scientific explanations and predictive models interchangeably, as elements of the same class. The question we are interested in is why models derived via the scientific method can extrapolate while models derived from the methods of deep learning over massive data cannot. More generally we are interested in what general prin-

ciples might allow any rule, model, or explanation to generalize to any degree. According to David Deutsch, one such principle is that models/explanations with large reach are "hard-to-vary" (HTV) while those with small reach are "easy-to-vary".[5] If true this conceptual dichotomy constitutes a profound insight, but to our knowledge it has not yet been formalized or applied to artificial intelligence.

3 Understanding the HTV principle

What makes Deutsch's HTV principle fascinating is that it was invented to support critical rationalism, the epistemology of Karl Popper. Deutsch has argued that a better understanding of Popperian epistemology is necessary before artificial general intelligence can be realized.[20] Popper famously rejected induction and believed induction is not required for the growth and development of scientific knowledge.[21] All of present day AI and deep learning is built on induction which is the process of starting with a blank slate and learning from data. To explain how science can work without induction Popper argued that theories can only be falsified by evidence, and not confirmed. In Popper's view theories are "bold conjectures" invented to solve problems and not learned directly from experience. To give a poignant example, the idea that stars are actually distant suns was a bold conjecture first proposed by Anaxagoras around 450 BC.

While experience can and does inform conjectures, in particular in the form of empirical tests, prior experience itself is theory-laden. In other words, observations cannot made in an unprejudiced or unbiased manner as Francis Bacon had prescribed.[22] To Popper, the question of which comes first, the theory or the observation, is much like the question of the chicken and the egg.[23] Scientific theories are built on previous scientific theories, which in turn were built on pre-scientific myths.[23] Thus, while empirical testing of theories plays a role in the form of falsifying certain theories while preserving others, Popper believes that fundamentally all theories originate "from within" rather being impressed into the mind from outside.[24] Interestingly, Popper believed critical rationalism was not limited to explaining the functioning of science but could be extended to advance psychology as well (ie. how the human brain works), although he admitted this was a bold and far-from-proven hypothesis.[24]

Deutsch's HTV principle helps solve an issue in critical rationalism, which is why we should reject explanations based on myths (ie explanations relying on the actions of Gods, demons, ghosts, etc) which are considered equally valid as theories alongside scientific theories as long as they are falsifiable. An inductivist would reject such myths on the grounds that there is no empirical evidence confirming them, but a critical rationalist cannot do that. The HTV principle asserts that we reject such explanations because they are easy to vary. In other words, they are "free floating" and not tied into previously established theories. Thus, if the data were different, they could easily adapt to that situation. An illustrative example which Deutsch elaborates in detail in his book *The Beginning of Infinity*[5] and in a 2009 TED talk[25] is the ancient Greek myth that the seasons were due to the periodic sadness of the god Demeter. This myth can

be easily varied by changing the gods and their behaviours, and thus is not a good explanation. The theory can also be varied internally, without changing the predictions it makes, for instance by substituting the gods employed with other ones, and can also be varied easily in the event that its predictions are invalidated. Both of these sources of variation make it a bad explanation, according to Deutsch.[5]

What makes a theory hard-to-vary? For Deutsch a key factor is constraints arising from the internal deductive logic of the theory. For instance, the modern day explanation of the seasons involves a series of geometrical deductions regarding the suns rays and the Earth's axis tilt. There are a few free parameters, such as the angle of tilt, but most the explanation is rooted in geometrical deductions which cannot be varied. Another important source of constraint is consistency with established knowledge, which we will return to in a bit.

4 How can we make the HTV principle more precise?

So far, the HTV principle has not been formalized. In this section we take tentative steps in the direction of formalization by exploring the relation of the HTV principle to two principles which have been formalized - the bias-variance trade-off and Occam's razor.

Superficially the HTV principle seems very related ot the bias-variance trade-off in classical statistics. Deep learning models with more parameters are more variable, but also more prone to overfitting their training data resulting in poor generalization to test data. However, recall we are interested in extrapolation, not classical generalization within the scope of the training distribution. Another important point is that the bias-variance trade-off has been shown to break down in machine learning as more parameters are added to the model - leading the the double descent curve, where beyond a certain threshold more parameters always helps and never hurts.[26,18,27] So, the bias-variance trade-off deals with a different issue and is also questionable on its own.

The HTV principle bares some resemblance to Occam's razor, a principle which was stated by William of Occam as "it is pointless to do with more what can be done with fewer"[28] and also as "Plurality (of entities) should not be posited without necessity".[29] Occam's razor is deeply embedded in the culture of science and is also quite popular amongst AI researchers. It is discussed in the most popular AI textbooks and the minimum description length principle[30] is an oft-cited formalization of Occam's razor which has been applied many areas of AI such as regularizing neural networks.[31] Solomonoff's theory of universal induction includes a constraint for Occam's razor,[32] formalized using Komologorov complexity as a measure of simplicity. Solomonoff's theory of induction was later used by Hutter to develop the AIXI model of universal artificial intelligence which has been influential in certain quarters of AI research.[33]

So is Deutsch's principle just a restatement of Occam's? HTV deals with variability while Occam's razor deals with the number of independent entities in the theory, so the answer depends on how one thinks these two are related. Models or

theories with more entities are appear to be more variable since the entities can be easily re-arranged to fit different data. Deutsch rejects the converse though - that simpler theories are less variable, saying that "there are plenty of simple explanations that are highly variable, such as 'Demeter did it"'.[5] Clearly the notion of "simplicity" plays a key role here.

5 Does the human brain run solely off of brute force fitting (induction) of easy-to-vary models?

Besides being a foundational question for neuroscience, answering this question has import for futurists and AI safety researchers looking to predict the development trajectory of AI and onset of superintelligence.[34] Taking a stand one way or the other on this issue is an example of what Armstrong and Sotala call a "metastatement" prediction about the entire field of AI, and differing stances on this question taken by Deutsch and others explains at least a bit of the high degree of variance between experts regarding the date superintelligent AI will be developed.[35] If the brain works entirely by fitting easy-to-vary models to big data, than reaching human level intelligence will be possible by scaling up deep learning with larger and larger models and datasets, suggesting a sooner time-to-arrival for superintelligent AI. However, if the generation of hard-to-vary explanations is an important part of human intelligence, the problem of how to program an AI to generate hard-to-vary explanations must be solved first before truly human-like AI can be produced. It is also possible though that the HTV principle might be learned though brute force fitting of highly variable deep learning models, as sort of an emergent principle which arises out of the process and then acts to constrain it. In that case the HTV principle would not be fundamentally important. To us this possibility seems unlikely, so we conjecture that to the extent the HTV principle is involved in human cognition it would have to be imbued genetically, having been stumbled on by evolution and not learned by induction of highly variable neuronal models on sense data. Still, we admit this is a possibility. Christopher Olah of OpenAI has posited something similar – in an interview published online he speculates that as deep learning models are scaled up they will eventually reach a point where they start to generate "crisp abstractions".[36] Given that currently many of the abstractions (features) being learned by today's overparametrized deep learning systems are very "uncrisp",[37] we find this also to be questionable, though.

Note though, that just as the way that airplanes fly is much different from birds, superintelligent AI might operate in a very different way from humans. So, while a positive answer to the question at hand suggests a sooner date for superintelligent AI, a negative answer is not as informative.

The brain has about 8.6×10^{10} neurons and an average of 10^3 connections (synapses) per neuron. Each synapse contains at least 4.7 bits of information.[38] Information is encoded in the brain both in the synaptic weights and the pattern of connections, which is arguably much more important.[39] About 80% of those neurons are in the cerebellum, which is responsible for motor control with

almost all the others in the cerebral cortex. Which neurons are connected to which is determined by a mix of genetic mechanisms which control the placement of connections during development and processes of dendritic grown and pruning during the lifetime of any given individual in response to its environment. The former can be viewed as an "outer-loop" optimization that occurs via evolution over many generations while the latter is part of an "inner-loop" optimization that occurs during an individual's lifetime. As a side note, Popper has noted that evolution, since it is based on random mutations which happen "from within" operates along the lines of his epistemology (conjecture and criticism/falsification).[24] The "inner loop" optimization, by contrast, appears to be largely based on classical induction from sense data. Both processes are known to occur although the relative importance of each is far from understood (the nature-vs-nurture debate).[40] The amount of data that can inform neural structure from genetics is limited however by size of the genome. This "genetic bottleneck" has been suggested to serve as a form of regularization on how evolution has shaped the brain.[40] The human genome contains about 750 megabytes of data total. About 2% of the genome is coding for proteins (15 Mb). While the degree of importance of the non-coding genome is largely unknown, if we assume 25% of the non-coding genome is used for regulatory purposes, that is about 200 Mb for the genetic bottleneck.

In what follows, we ignore the compex question of the importance of how much is learned from genetics vs environment. Instead we focus on the question of how far brute-force fitting to sense data alone can go. Conventional thinking in neuroscience has been that the brain cannot operate on brute-force fitting of sense data alone, because the amount of data provided is too scant to train the number of parameters involved. This "poverty of the stimulus" was the key motivation for Chomsky to propose a genetic component to human grammar which he called "universal grammar".[41] However, recently Hasson et al. have argued that the amount of data streaming in though the senses is enough to allow for brute-force fitting.[17] Quantifying the amount of information streaming into the brain through the senses is very difficult, but it may be quite large – Zimmerman estimates that the channel capacity of the human visual system is 10^7 bits/s while the channel capacity of the auditory system is 10^5 bits/s.[42] The actual amount of information is much lower due to redundancy effects, however if we assume it is 100 times lower than learning the $4x10^{14}$ bits necessary to specify all $9x10^{13}$ synapses in the brain is theoretically possible in 133 years. Regardless of the relevance of this back-of-the envelope calculation, it is safe to say that the brain does process very large amounts of data. For instance a child has seen many faces by an early age under a variety of angles, distances, and lighting conditions. As Hasson points out, humans are subject to an "other-race" effect whereby they find faces of different ethnicities more difficult to distinguish, suggesting a type of brute force fitting which struggles to extrapolate even slightly.[17] A study based on audio recording in a child's home suggests that children hear at least 5-10 million words per year.[43] By age 10, this suggests 100 million words may

have been heard. This is very large but still 3000 times smaller than the natural language model GPT3's training corpus, which contained 300 billion tokens.[4]

Thus, it seems that a large part of the brain, at least that dedicated to raw perception of objects and possibly a large part of language ability is learned by fitting easy-to-vary models to sensory data. However even Uri Hasson, who argued for this position recently,[17] also believes that have additional abilities which go beyond this, pointing to human's ability learn mathematics such as calculus and also develop physical theories which can extrapolate.[44]

What about "common sense" knowledge, or "background knowledge"? Some of common sense knowledge comes in the form of "rules of thumb", which are rules extracted from experience through induction. They are brute facts, often "free floating" and unconnected to other facts. Suppose we are talking to someone who has no knowledge of how the human body works. They still may be able to treat certain diseases using a long list of rules of thumb gained from experience. For instance:

1. Aspirin helps with headaches. If Aspirin doesn't work, try ibuprofen next.
2. Drinking lots of water helps ameliorate cold symptoms.
3. Homeopathic remedies do not work. (note one can mistakenly go the wrong way here!)

The person wouldn't be able to explain any of these facts but instead would justify them by pointing to how often they have been confirmed by prior experience. Explanatory theories, on the other hand, are required whenever trying to predict something you haven't seen before. Consider the following questions:

- How many marshmallows can fit on a toothpick?
- What would happen if you poured a bottle of bleach into the fuel tank of a car?
- What would happen if the United States cut off all trade with France tomorrow?
- What would happen if you deluted hot sauce 10 times and repeated this 10x dilution process 10 times?

We must use some form of common sense reasoning to answer these questions, reasoning which uses explanatory theories. The GPT2 language model (which is a type of brute-force fit, highly-variable model) is very bad at answering these sorts of questions[45]. GPT3 does much better, but still stumbles in this area.[46]

To conclude, the human capacity to build both common sense explanatory theories and scientific explanatory theories is where the HTV principle may be important.

6 Conclusion

Present day AI is built on top of two philosophical principles - induction and Occam's razor. Popper and Deutsch reject induction, arguing for critical rationalism instead, which is based on conjecture, criticism, and falsification. Deutsch

also rejects Occam's razor, calling it a "misconception".[5] The HTV principle was introduced by Deutsch to improve critical rationalist epistemology. However, the HTV principle may be a useful principle to guide further AI research regardless of which view on epistemology one ultimately deems correct. In this work we took some first steps towards exploring the relevance of this principle to artificial intelligence and neuroscience. The brain appears to operate largely on brute-force fitting of easy-to-vary models. At the same time however, humans possess additional capacities for reasoning (ie Kahneman's "system 2") and are capable of inventing explanations (both of the common sense and scientific variety) which are capable of extrapolation. Both of these seem impossible to obtain with brute-force fitting of highly variable models alone. The HTV principle is very similar to Occam's razor, but distinct if one accepts that simple theories can be highly variable. The precise relationship between variability and simplicity requires further study. Even if the HTV principle ultimately reduces to Occam's we believe it may be a useful frame. Further study of this subject will be the focus of future work. We believe work in this area may shed light on a key issue facing AI research today - how to build systems that can generate "good explanations" of the world and therefore can extrapolate to new situations.

Funding & disclaimer

No funding sources were used in the creation of this work. The author (Dr. Daniel C. Elton) wrote this article in his personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

Acknowledgements

The author appreciates helpful discussions with Dr. Felix Faber on this subject and feedback from Bruce Nielson, who has written several blog posts on the HTV principle and critical rationalism on Medium.

References

1. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, December 2015.
2. Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, October 2019.

3. David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.
4. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv e-prints: 2005.14165*, 2020.
5. David Deutsch. *The Beginning of Infinity: Explanations That Transform the World*. Viking Adult, 2011.
6. D. Hackethal. *A Window on Intelligence: The Philosophy of People, Software, and Evolution - and Its Implications*. 2020.
7. David Davenport. Explaining everything. In *Fundamental Issues of Artificial Intelligence*, pages 341–354. Springer International Publishing, 2016.
8. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
9. Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
10. Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, April 2020.
11. Tom Simonite. When it comes to gorillas, google photos remains blind. *WIRED*, Nov 2018.
12. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.
13. Ken Kansky, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, D. Scott Phoenix, and Dileep George. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1809–1818. PMLR, 2017.
14. Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R.

Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, September 2019.
15. W. Patrick Walters and Mark Murcko. Assessing the impact of generative AI on medicinal chemistry. *Nature Biotechnology*, 38(2):143–145, January 2020.
16. Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, 2019.
17. Uri Hasson, Samuel A. Nastase, and Ariel Goldstein. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, February 2020.
18. Daniel C. Elton. Self-explaining AI as an alternative to interpretable AI. In *Artificial General Intelligence*, pages 95–106. Springer International Publishing, 2020.
19. Frank Watson Dyson, Arthur Stanley Eddington, and C. Davidson. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 220(571-581):291–333, January 1920.
20. David Deutsch. Creative blocks. *Aeon*, October 2012.
21. K.R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & K. Paul, 1963.
22. Francis Bacon. *Novum Organum*. 1620.
23. K.R. Popper. *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, 1979.
24. K.R. Popper, K.R. Popper, and M.A. Notturno. *The Myth of the Framework: In Defence of Science and Rationality*. In Defence of Science and Rationality. Routledge, 1996.
25. David Deutsch. A new way to explain explanation. *TED Talk*, 2009.
26. Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, July 2019.
27. Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv eprints: 1912.02292*, 2019.
28. William of Ockham. *Summa totius logicae*. 1323.
29. J. Badius and J. Trechsel. *Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi: Centilogium theologicum*. Johannes Trechsel, 1495.
30. J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, September 1978.
31. Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory - COLT '93*. ACM Press, 1993.
32. Samuel Rathmanner and Marcus Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136, June 2011.
33. Marcus Hutter. A theory of universal artificial intelligence based on algorithmic complexity. *arXiv e-prints: cs/0004001*, 2000.
34. N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

35. Stuart Armstrong and Kaj Sotala. How we're predicting AI – or failing to. In *Topics in Intelligent Engineering and Informatics*, pages 11–29. Springer International Publishing, 2015.
36. Evan Hubinger. Chris Olah's views on AGI safety. *https://www.lesswrong.com/posts/X2i9dQQK3gETCyqh2/chris-olah-s-views-on-agi-safet*, 2020.
37. Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 125–136, 2019.
38. Thomas M Bartol, Cailey Bromer, Justin Kinney, Michael A Chirillo, Jennifer N Bourne, Kristen M Harris, and Terrence J Sejnowski. Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife*, 2015.
39. S. Seung. *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt, 2012.
40. Anthony M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1), August 2019.
41. N. Chomsky. *Aspects of the Theory of Syntax*. Aspects of the Theory of Syntax. MIT Press, 2014.
42. M. Zimmermann. Neurophysiology of sensory systems. In *Fundamentals of Sensory Physiology*, pages 68–116. Springer Berlin Heidelberg, 1986.
43. Brandon C. Roy, Michael C. Frank, Philip DeCamp, Matthew Miller, and Deb Roy. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668, September 2015.
44. Karl Polich. "Robust Fit to Nature" - interview with Uri Hasson. *https://dataskeptic.com/blog/episodes/2020/robust-fit-to-nature*, 2020.
45. Gary Marcus. GPT-2 and the nature of intelligence. *The Gradient*, 2020.
46. Alyssa Vance. Fun with GPT-3. *https://rationalconspiracy.com/2020/07/31/fun-with-gpt-3/*, 2020.