https://doi.org/10.1038/s42254-022-00538-z

Should artificial intelligence be interpretable to humans?

Matthew D. Schwartz

As artificial intelligence (AI) makes increasingly impressive contributions to science, scientists increasingly want to understand how AI reaches its conclusions. Matthew D. Schwartz discusses what it means to understand AI and whether such a goal is achievable — or even needed.

It is difficult to deny the permanent role AI is establishing in the modern scientific enterprise. Yet, even when AI makes testably correct predictions, practitioners and critics are wary of trusting it unless its success can be explained in simple terms. This desire for understanding is instinctive, but as time goes on it will be harder and harder to satisfy. With a little perspective it quickly becomes clear that our quest to understand AI's understanding is short-sighted and reflective of our anthropocentric bias.

The evolution of biological and artificial intelligence

Giving up the hope for understanding in human terms is not so different from other ways in which science has forced us to accept our insignificance. Astronomy taught us that we are not in the centre of the Solar System. Geology taught us that our planet is not six thousand years old. String theory suggests that our Universe is not the unique solution to a theory of everything. Today we can lament the limitations of our own biology. A valuable perspective is provided in a prescient article by Freeman Dyson from 1979[1]. Dyson observed that it takes about 10^6 years to evolve a species, 10^8 years to evolve a genus and "less than 10^{10} years to evolve all the way from the primaeval slime to *Homo sapiens*." We think of our phylogenetic tree as evolving downward, trunk to leaf, but Dyson saw it evolving upwards. After 10^{11} years we will need an extension above kingdom, after 10^{12} years a clade above that, and so on. After, say, 10^{20} years, the distinction between us and protozoa will be a rounding error. From this perspective, it is hard to imagine how anything about our species can be special at all.

Where do current machines fall in the biological spectrum? The human brain has around 100 trillion synapses, compared to say a cat with around 10 trillion synapses or a honeybee with of order 1 billion synapses. In contrast, consider Google's large language model Pathways Language Model (PaLM) with 540 billion parameters[2]. Just a few years ago, the state of the art for similar models was 100 million parameters. So, by a crude estimate, these models grow by around a factor of 10 per year (Fig. 1). Contrast this with the humanoid brain which took 10 million years to grow by a factor of 3. In other words, PaLM is currently somewhere in Chordata, and large language models will reach *Homo sapiens* within a few years and move on beyond us by the end of the decade.

How machines learn science

What can a model like PaLM do now? An impressive application is Google's Minerva project[3]. Minerva took PaLM and fine-tuned it to learn a new language: LaTeX. Crawling the Internet, it studied any website where science is mentioned or explained. The result was a machine that could not only answer an impressive variety of high-school and college-level science questions, but also explain its answers with text and equations. Its effectiveness was enhanced with few-shot learning and chain-of-thought prompting[4]. That is, it shows its work. Although one can validate that Minerva is (often) correct, one may question whether Minerva understands the science it is explaining. Perhaps it is just generalizing from similar examples it has seen and memorized. But is that really so different from what human students do?

One might argue that there is less scrapable data available for science beyond college, so continued progress will slow down. However, it is inconceivable that we have already reached the endpoint of machine evolution. Networks will continue to grow. Of course, size isn't everything. It takes several months for the 1,024 FLOPs needed to train PaLM and requires around 1 million kWh of energy. In contrast, it takes around 16 years for a human to solve the same problems as Minerva and around 10,000 kWh of energy consumption. But the efficiency of training has also been improving considerably, and energy efficiency seems to follow an analogue of Moore's law. Coupled with the ability of machines to teach themselves (such as AlphaGo), it should be possible for them to generate their own training data. Machines have a passable understanding of college-level science now. With the current rate of progress, how long before they will be generating original arXiv papers that put the work of human scientists to shame?

Limits to interpretability

Both biological and artificial intelligence seem to evolve exponentially, but with exponents that differ by a factor of a million. In this sobering context, we can explore current approaches to interrogating the machines to probe their understanding. To use a concrete example, consider the problem from high-energy physics of identifying unstable particles such as top quarks from their decay products[5]. A traditional approach would use domain knowledge to characterize these decay products, such as looking for evidence of a W boson in the particle shower. A machine-learning approach would be to train a neural network on simulated data without any insight from the standard model. The machines perform significantly better than the traditional approach. But why? Do the machines understand the physics better than we do? To answer these questions, one can explore the latent space directly, fit the neural network output to symbolic forms, or use Shapley values or other techniques[6]. The implicit hope is that a selection of interpretable observables, such as evidence for the W boson within the decay products, could work as well as the machine if only we could combine them more effectively. But this approach largely misses the point. The machine has a qualitatively different way of

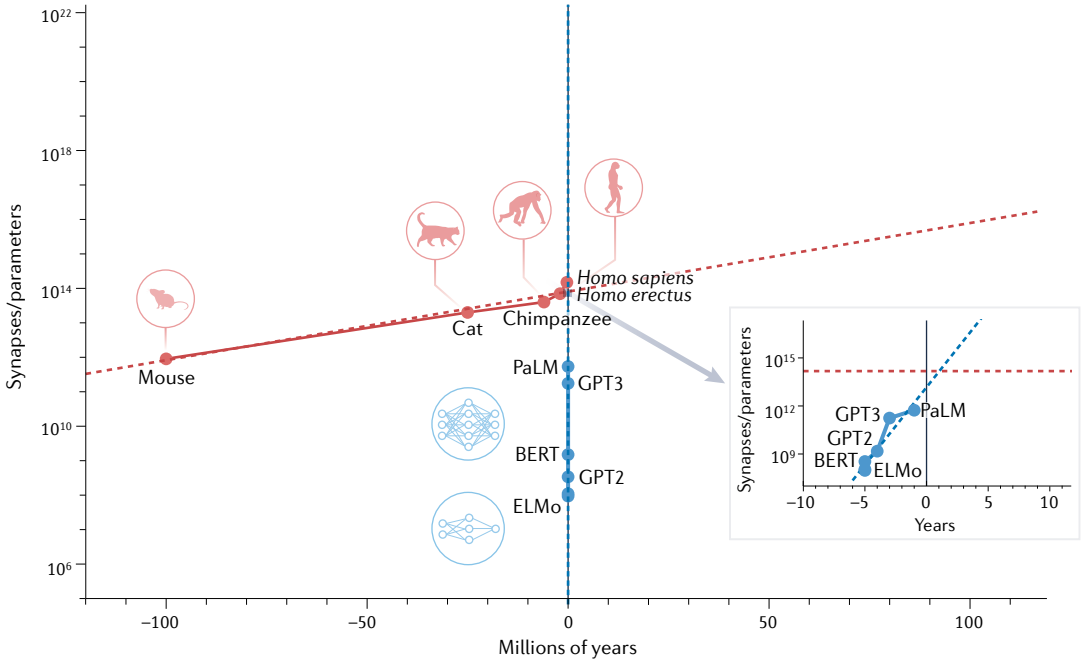

Fig. 1 | The evolution of biological and artificial intelligence takes place on dramatically different timescales. Any hope of interpreting and understanding AI will exponentially fade. Some example data points are highlighted in the evolution of biological (red) and artificial (blue) intelligence. The dashed lines represent the linear regression to these points. The acronyms in the figure are: Pathways Language Model (PaLM), Embeddings from Language Model (ELMo), Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT).

understanding data than we do. Even if a handful of simple observables can be extracted from the machine, the machine always does better than these observables alone. It is exactly this marginal enhancement that is key to its power, not the simple observables to which we reduce it. The gap between the interpretable component of the machines success and its full power will only increase with time. All we can ever do is reduce the machine's understanding to our language since we cannot process its language directly.

Another concept commonly associated with interpretability is symmetry. If a machine can learn a symmetry, then surely it understands the physics. In a sense, symmetry is important not because the laws are ineffective without it (electromagnetic fields behave the same whether or not we know about relativity), but because they make it simpler for us to understand the physics. We like to chunk our models into manageable blocks, master these, then generalize. The simpler the block, the easier it is for us to see its relation to deeper, more powerful concepts. Since symmetries are relatively uncommon, we have developed other tools to assist our cognition, such as notation. In physics, good notation is of incalculable value. Examples include the Einstein summation convention, the bra-ket notation of Dirac, or Feynman diagrams. As we should not expect the internal notation of a machine to resemble ours, we should also probably not expect it to value symmetries as much as we do. There may be concepts even more powerful than symmetries that machines can appreciate, yet we cannot. These concepts may never be understandable to human beings.

Beyond our horizon

In a Perspective[7] in this issue, Mario Krenn et al. argue that if a machine understands something, it should be able to transfer this knowledge to a human being. This is along the lines of an idea often attributed to Ernest Rutherford, that you do not understand a scientific concept unless you can explain it to a child. But a Nobel Laureate and a child have identical cognitive capacity on evolutionary timescales. A better question is how are we to transfer knowledge to an intellect that perceives the world in a qualitatively different way than us, like Wittgenstein's lion ("If a lion could talk, we would not understand him") or Nagel's bat[8]? We should not be expected to teach calculus to a cat, or painting to a protozoon. So why should we expect machines to be able to explain things to us? We are just a blip on the infinite continuum of cognition, and our method of understanding is not unique. But this is a good thing. Much as we can admire athletes, artists, or scientists who can achieve what is beyond us, we can admire machines for possessing understanding that is outside of our umwelt. I personally cannot wait to read research papers generated by AI, solving problems with which humanity has long struggled, and to witness the quality of that research advance beyond my cognitive horizon.

Matthew D. Schwartz ✉
Department of Physics, Harvard University, Cambridge, MA, USA.
✉e-mail: schwartz@g.harvard.edu

Published online: 02 November 2022

References

1. Dyson, F. J. Time without end: Physics and biology in an open universe. *Rev. Mod. Phys.* **51**, 447 (1979).
2. Chowdhery, A. et al. PaLM: Scaling language modeling with pathways. Preprint at https://doi.org/10.48550/arXiv.2204.02311 (2022).
3. Lewkowycz, A. Solving quantitative reasoning problems with Language models. Preprint at https://doi.org/10.48550/arXiv.2206.14858 (2022).
4. Wei, J. et al. Chain of thought prompting elicits reasoning in large language models. Preprint at https://doi.org/10.48550/arXiv.2201.11903 (2022).
5. Schwartz, M. D. Modern machine learning and particle physics. *Harvard Data Sci. Rev.* https://doi.org/10.1162/99608f92.beeb1183 (2021).
6. Grojean, C. et al. Lessons on interpretable machine learning from particle physics. *Nat. Rev. Phys.* **4**, 284–286 (2022).
7. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* https://doi.org/10.1038/s42254-022-00518-3 (2022).
8. Nagel, T. What is it like to be a bat? *Philos. Rev.* **83**, 435–450 (1974).

Competing interests

The author declares no competing interests.