https://doi.org/10.1038/s41567-024-02403-5

Artificial intelligence needs a scientific method-driven reset

By Luís A. Nunes Amaral

AI needs to develop more solid assumptions, falsifiable hypotheses, and rigorous experimentation.

Recently, artificial intelligence (AI) has been discussed as either the solution to all human challenges — human cognition, aging, climate change — or the technology that will finally lead to the demise of humanity. Often overlooked is the fact that not only are these challenges characterized by complexity, but so is the development of machine intelligence. Complex systems are highly adaptable to a broad range of conditions and reward systems. Ecosystems can resist the extinction of numerous species. Our heart rate is not set by a single, highly accurate 'clock cell', but by many interacting clocks of dubious reliability.

Consider the AI approaches to computer vision. The current framework had its foundation in the idea of trying to replicate the learning processes in the human brain. This hypothesis first yielded the percepton: a single layer of 'neurons', each separately weighing the importance of each pixel, and then contributing to the collective identification of the object in the image.

Although perceptons successfully recognized digits, they failed when they were exposed to more challenging tasks. The field's solution was adding internal layers of neurons, marking the birth of deep learning. But unlike the neuronal networks in the human brain, deep learning requires vast amounts of energy and data. Indeed, the success of deep learning originated in the confluence of vast amounts of money and data at a few tech giants, which enabled access to nearly unlimited computational resources for model training. Driven by tech's reward system, the allocation of such immense resources often resulted in astonishing, but poorly validated, claims.

Unsurprisingly, the access to ever-growing financial and technological resources reduced the need to understand how AI architectures actually 'learn' (that is, estimate parameter values), with efforts focused on accuracy improvements. Similarly, the abundance of available data hid the extent and impact of biases on poor or non-white people. Add to this sloppy[1] data labelling, and you end up with a technology that consumes vast amounts of resources[2], works in a biased manner[3], and is easily deceived[4].

Looking in from the outside, it has become clear that AI research is in dire need of a makeover of its goals, metrics of success and validation methods. Ideally, such a makeover must be guided by the scientific method, thus relying on prior knowledge, falsifiable hypotheses, and rigorous experimentation. Prior knowledge is what we believe to be true a priori. Deep learning assumes that nodes learn like neurons, and that neural networks learn like the brain. Neither of these critical assumptions has been tested with any rigour.

A hypothesis is falsifiable if one can conceive of tests that can disprove it. However, many of the tests that AI applications are put through are unreliable. For example, in many computer vision tasks, a model is determined to have provided a correct answer if any of its top five guesses matches a human-assigned label. But it was only recently that Tsipras et al.[1] reported that serious annotation issues affect a large fraction of ImageNet — a popular dataset for vision recognition tasks. These issues raise questions about whether aiming for nearly perfect prediction accuracy in benchmark data sets is truly indicative of the desired performance in the actual task.

The sloppiness has been in full display in the context of uber-hyped large language models such as ChatGPT. Many press releases and preprint submissions have claimed extraordinary performance in a plethora of human cognitive tasks. Earlier this year, a preprint appeared on arXiv reporting that GPT-4 could achieve a perfect solve rate on a test corpus of questions from problem sets and exams from across various MIT courses. Shortly after the announcement, a trio of MIT undergraduates exposed a staggering number of ethical, methodological, and technical problems in the paper's claims[5], leading to its withdrawal.

AI research has featured decades of hype and crash cycles. To prompt hype periods and the concomitant financial rewards, researchers and institution have often resorted to unrealistic claims. For instance, in the proposal for the 1956 Dartmouth Summer Research Project on Artificial Intelligence, the organizers stated that significant advances in one or more fundamental AI tasks, like language processing, abstract thinking and other common human tasks, could be achieved "if a carefully selected group of scientists work on it together for a summer"[6]. However, initial claims on the ability of machines to translate between Russian and English were quickly debunked, prompting a drastic reduction in governmental funding for the field.

I believe that use of the scientific method (and Occam's razor) can improve the situation. Again, taking computer vision as a paradigmatic example, we need to truly understand how the model parameters are learned. Is the limiting factor the size and depth of the internal network, or is it the use of unsophisticated optimization algorithms? We also need to clarify the role of task complexity in the generalizability of model training — discriminating auto vehicles from mammals or individual cows in a herd are unlikely to be accomplished by the same model.

Finally, we need to validate model performance in cases for which we have easy access to the correct labels. This is where the systematic development of synthetic datasets is critical. All these steps have been crucial to solving other important problems in computer science — such as finding modules within large complex networks — to which physicists contributed in a significant manner.

Luís A. Nunes Amaral
School of Engineering, Northwestern University, Evanston, IL, USA.
e-mail: amaral@northwestern.edu

Published online: 23 February 2024

References

1. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A. & Madry, A. In *International Conference on Machine Learning* 9625–9635 (PMLR, 2020).
2. Strubell, E., Ganesh, A. & McCallum, A. *AAAI* **34**, 13693–13696 (2019).

3. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
4. Goodfellow, I. J., Shlens, J. & Szegedy C. In *3rd International Conference on Learning Representations* (ICLR, 2015).
5. Chowdhuri, R., Deshmukh, N. & Koplow, D. *No, GPT4 can't ace MIT*. https://flower-nutria-41d.notion.site/No-GPT4-can-t-ace-MIT-b27e6796ab5a48368127a98216c76864 (2023).
6. McCarthy, J., Minsky, M., Rochester, N. & Shannon, C. E. *AI Mag.* **27**, 12–14 (2006).

Competing interests

The author declares no competing interests.