AI for science: the easy and hard problems

Sam Gershman
Harvard University

"If we knew what we were doing, it would not be called research, would it?"- Albert Einstein

The easy problem

Most work on applying AI to science has focused on what might be called the "easy problem" of AI science (this is a relative term, since the easy problem is actually quite hard). A scientist specifies a function that they want to optimize (e.g., a function that scores protein folding structure or fusion reactor designs). AI optimization tools can then be applied to this problem. So far, this kind of application has been highly successful (e.g., AlphaFold[1]).

What makes this problem "easy" is not the form of the solution (which may require a great deal of engineering work) but rather the form of the problem. It is clear from the beginning what needs to be optimized, and what kinds of tools can be brought to bear on this problem. The engineering breakthrough comes from building much better versions of these tools. In other words, the problem is relatively easy because it does not require any *conceptual* breakthroughs of the sort involved in the discovery of relativity theory, genetics, or the periodic table.

Are these conceptual breakthroughs just patterns that can be discovered with a sufficiently powerful pattern recognition system? In a sense yes, but before that can happen, something has to tell the pattern recognition system what kind of patterns are interesting, important, and useful. What problem is the pattern recognition system designed to solve, and where does this come from?

The hard problem

The fundamental barrier to automating science is conceptual. Great scientists aren't simply extraordinary optimizers of ordinary optimization problems. It's not like Einstein had a better function approximator in his brain than his peers did, or Mendeleev had a better version of backprop in his brain. More commonly, great scientists are ordinary optimizers of extraordinary optimization problems. It is the formulation of the problem, not its solution, that is the truly hard problem.

One might be tempted to relegate the hard problem to the fringes of "revolutionary science" (in Kuhn's sense[2]), which rarely erupt into mainstream scientific practice, whereas the easy problem

[1] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583-589.
[2] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press..

occupies the focus of the “normal science” that scientists spend most of their time on. However, normal science is *not* simply optimization. This is obvious to any first-year grad student trying to figure out what to work on. Normal science isn’t a catalog of optimization problems waiting to be solved by a queue of grad students. The fundamental barrier for grad students is the same one facing AI scientists: it is the conceptual problem of formulating an optimization problem. This encompasses both major conceptual breakthroughs, like relativity theory, and the more modest ones achieved by grad students on a regular basis, which nonetheless remain out of reach for existing AI systems.

Looking backward

Much of the classic work on AI science (mainly by Simon, Langley, and their collaborators,[3] but also more recently by Schmidt & Lipson,[4] Udrescu & Tegmark,[5] and others) focused on the easy problem. For Simon and Langley, this approach was premised on the psychological thesis that scientific cognition was essentially the same as regular problem solving, only applied to a different (and sometimes more challenging) set of problems. Consequently, they developed algorithms that emulated human problem solving, and applied these to scientific discovery. This approach was criticized by Chalmers, French, & Hofstadter[6] because it endowed the algorithms with a representation of the problem that already had the basic primitives needed for the final theory. In other words, it skirted the problem of representation: where do the primitives come from, and how do we know if we have the right ones? Simon insisted (contra Popper) that there was a logic of scientific discovery, but his was in fact a logic of scientific problem *solving* (i.e., optimization), not discovery in the sense of problem *creation*. The latter involves representation learning, but also something deeper, as I argue below.

Moving forward

In contemplating how to build AI systems that solve the hard problem, it is instructive to look at how human scientists do it. At a high-level, human scientists break this into several sub-problems:

- *Domain specification.* What are the relevant phenomena that need to be explained by a theory?
- *Constraint specification*. What kinds of constraints need to be imposed on a theory based on existing knowledge (both domain-specific and domain-general)?

[3] Bradshaw, G. F., Langley, P. W., & Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science, 222*, 971-975.
[4] Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*, 81-85.
[5] Udrescu, S. M., & Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances, 6*, 2631.
[6] Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence, 4*, 185-211.

Once the domain and constraints have been specified, we can define an optimization problem (theory search); hence, we have converted the hard problem into the easy problem. However, it is uncommon to do a single pass from hard to easy, because scientists often realize that the problem they're solving is the wrong one. This may happen for several reasons. One is the realization that a theory is internally inconsistent or paradoxical. Another is the realization that the theory may (with suitable modification) be able to explain a broader range of phenomena, prompting a respecification of the domain. Conversely, phenomena which were previously included in a domain may need to be excluded if no adequate unifying theory is found for all the phenomena. Respecification can also happen when new empirical phenomena are reported. In a related vein, constraint respecification can happen when domains are merged, split, expanded, or shrunk. The key point is that problem creation and problem solving are cyclically coupled in scientific practice.

An important and elusive feature of problem creation is that *it is not a data modeling problem*. The selection of what to model and and what constraints to condition on are antecedent to any data modeling problem. It is also not reducible to a representation learning problem, in the sense of figuring out how raw sensory input maps to abstract representations. Of course, that problem also needs to be solved, but first the scientist needs to know what problems the representations are being used to solve.

Sociological, aesthetic, and utility considerations enter at the problem creation stage. Building an AI scientist is as much about shaping its tastes, style, and preferences as it is about endowing it with powerful problem-solving abilities. Again, a look at how we train human scientists is instructive: a good graduate advisor educates students about what problems matter, what phenomena are interesting, which explanations count, and so on. These considerations can't be brushed aside as subjective factors irrelevant to the purely technical problems facing AI systems; they are in fact constitutive of those technical problems. Without them, the technical problems would not exist.

A research program for attacking the hard problem should begin with the cognitive science of science,[7] focusing on the understudied subjective, creative aspects discussed above and how they interact with the objective aspects of problem solving. Once we understand what human scientists are doing with enough precision that we can formalize these aspects, we can try to leverage these insights to build scalable AI scientists. At least initially, it is unlikely that these will be standalone systems, but rather more like research assistants or first-year grad students: curious agents with some technical competence but in need of expert guidance. This guidance can come in the form of natural language instruction, reading curricula, and demonstrations.

[7] For an introduction, see Thagard, P. (2012). *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*. MIT Press.