More Is Different for AI

JAN 4, 2022 • 3 MIN READ

Machine learning is touching increasingly many aspects of our society, and its effect will only continue to grow. Given this, I and many others care about risks from future ML systems and how to mitigate them.

When thinking about safety risks from ML, there are two common approaches, which I'll call the **Engineering** approach and the **Philosophy** approach:

- The Engineering approach tends to be empirically-driven, drawing experience from existing or past ML systems and looking at issues that either: (1) are already major problems, or (2) are minor problems, but can be expected to get worse in the future. Engineering tends to be bottom-up and tends to be both in touch with and anchored on current state-of-the-art systems.
- The Philosophy approach tends to think more about the limit of very advanced systems. It is willing to entertain thought experiments that would be implausible with current state-of-the-art systems (such as Nick Bostrom's paperclip maximizer) and is open to considering abstractions without knowing many details. It often sounds more "sci-fi like" and more like philosophy than like computer science. It draws some inspiration from current ML systems, but often only in broad strokes.

I'll discuss these approaches mainly in the context of ML safety, but the same distinction applies in other areas. For instance, an Engineering approach to AI + Law might focus on how to regulate self-driving cars, while Philosophy might ask whether using AI in judicial decision-making could undermine liberal democracy.

While Engineering and Philosophy agree on some things, for the most part they make wildly different predictions both about what the key safety risks from ML will be and how we should address them:

- Both Engineering and Philosophy would agree on some high-level points: they would agree that misaligned objectives are an important problem with ML systems that is likely to get worse. Engineering believes this because of examples like the Facebook recommender system, while Philosophy believes this based on conceptual arguments like those in *Superintelligence*. Philosophy is more confident that misaligned objectives are a big problem and thinks they could pose an existential threat to humanity if not addressed.
- Engineering and Philosophy would both agree that out-of-distribution robustness is an important issue. However, Philosophy might view most engineering-robustness problems (such as those faced by self-driving cars) as temporary issues that will get fixed once we train on more data. Philosophy is more worried about whether systems can generalize from settings where humans can provide data, to settings where they cannot provide data even in principle.
- Engineering tends to focus on tasks where current ML systems don't work well, weighted by their impact and representativeness. Philosophy focuses on tasks that have a certain abstract property that seems important, such as imitative deception.

In my experience, people who strongly subscribe to the Engineering worldview tend to think of Philosophy as fundamentally confused and ungrounded, while those who strongly subscribe to Philosophy think of most Engineering work as misguided and orthogonal (at best) to the long-term safety of ML. Given this sharp contrast and the importance of the problem, I've thought a lot about which—if either—is the "right" approach.

Coming in, I was mostly on the Engineering side, although I had more sympathy for Philosophy than the median ML researcher (who has ~0% sympathy for Philosophy). However, I now feel that:

- **Philosophy is significantly underrated by most ML researchers**.
- The Engineering worldview, taken seriously, actually implies assigning significant weight to thought experiments.

On the other hand, I also feel that:

- Philosophy continues to significantly underrate the value of empirical data.
- Neither of these approaches is satisfying and we actually have **no single good approach** to thinking about risks from future ML systems.

I've reached these conclusions through a combination of thinking, discussing with others, and observing empirical developments in ML since 2011 (when I entered the field). I've distilled my thoughts into a series of blog posts, where I'll argue that:

1. *Future ML Systems Will be Qualitatively Different* from those we see today. Indeed, ML systems have historically exhibited qualitative changes as a result of increasing their scale. This is an instance of "More Is Different", which is commonplace in other fields such as physics, biology, and economics (see *Appendix: More Is Different in Other Domains*). Consequently, we should expect ML to exhibit more qualitative changes as it scales up in the future.
2. Most discussions of ML failures are anchored either on existing systems or on humans. *Thought Experiments Provide a Third Anchor*, and having three anchors is much better than having two, but each has its own weaknesses.
3. If we take thought experiments seriously, we end up predicting that *ML Systems Will Have Weird Failure Modes*. Some important failure modes of ML systems will not be present in any existing systems, and might manifest quickly enough that we can't safely wait for them to occur before addressing them.
4. My biggest disagreement with the Philosophy view is that I think *Empirical Findings Generalize Surprisingly Far*, meaning that well-chosen experiments on current systems can tell us a lot about future systems.

This post is the introduction to the series. I'll post the next part each Tuesday, and update this page with links once the post is up. In the meantime, leave comments with any thoughts you have, or contact me if you'd like to preview the upcoming posts and leave feedback.

Jacob Steinhardt

3 Comments

Sign in to join the conversation.

Mikhail Grankin 11 months ago

"However, Philosophy might view most engineering-robustness problems (such as those faced by self-driving cars) as temporary issues that will get fixed once we train on more data. "

I believe this is a typo. Philosophy->Engineering

♡ 0

Phil Goetz 6 months ago

I think both Engineering and Philosophy would think that. I interpreted Jacob's statement to mean that "philosophers" would see such robustness problems as less-important than engineers do.

♡ 0

Phil Goetz 6 months ago

How does the division of a science into "engineering" vs. "philosophy" differ from engineering vs. research, or engineering vs. theory?

I like your idea of phrasing what I think of as theory as philosophy, because it makes it a little more clear that philosophy isn't dead; it's moved into the sciences.

But I have a difficulty with your dichotomy, because I'm in the habit of calling rationalist philosophy "philosophy", and empiricist philosophy "science". What you're calling ML philosophy has a rationalist attitude (reasoning about things unseen), yet must have enough empirical grounding to not spin off into BS the way rationalist philosophy always does. For instance, it's still got to be nominalist, quantitative, and conceptualize its models as ranging over continuums, not integers.

... which Superintelligence doesn't do beyond being nominalist. It really is just dialectic. Hmm. I'm struggling here, because to admit dialectic can work would really be a damning indictment of rationalist philosophers. It seems more a priori likely, and more kind, to say that dialectic just can't work except by chance, than to say that it can, but nearly all philosophers since Epicurus have been morons.

Although, arguably, everyone since Kant who was intelligent enough to do philosophy, realized they should go into science instead. I could make up other stories as well. The salons of Europe turned philosophy into entertainment. The French Revolution killed off most of continental Europe's intellectuals. Romanticism poisoned the brains of continental Europe. The easy accessibility of so many narratives makes me suspect there are too many degrees of freedom, and so any attempt to explain the unreasonable ineffectiveness of philosophy is doomed to overfit the data.

Edited by the author on 3/6/2023

♡ 0

← Previous Post

Next Post →