Bounded Regret

Thought Experiments Provide a Third Anchor

JAN 18, 2022 • 4 MIN READ

Previously, I argued that we should expect future ML systems to often exhibit "emergent" behavior, where they acquire new capabilities that were not explicitly designed or intended, simply as a result of scaling. This was a special case of a general phenomenon in the physical sciences called More Is Different.

I care about this because I think AI will have a huge impact on society, and I want to forecast what future systems will be like so that I can steer things to be better. To that end, I find More Is Different to be troubling and disorienting. I'm inclined to forecast the future by looking at existing trends and asking what will happen if they continue, but we should instead expect new qualitative behaviors to arise all the time that are not an extrapolation of previous trends.

Given this, how can we predict what future systems will look like? For this, I find it helpful to think in terms of "anchors"---reference classes that are broadly analogous to future ML systems, which we can then use to make predictions.

The most obvious reference class for future ML systems is current ML systems---I'll call this the **current ML** anchor. I think this is indeed a pretty good starting point, but we've already seen that it fails to account for emergent capabilities.

What other anchors can we use? One intuitive approach would be to look for things that humans are good at but that current ML systems are bad at. This would include:

- Mastery of external tools (e.g. calculators, search engines, software, programming)
- Very efficient learning (e.g. reading a textbook once to learn a new subject)
- Long-term planning (e.g. being able to successfully achieve goa

Models sufficiently far in the future will presumably have these sorts of capabilities. While this still leaves unknowns---for instance, we don't know how rapidly these capabilities will appear---it's still a useful complement to the current ML anchor. I'll call this the **human anchor**.

A problem with the human anchor is that it risks anthropomorphising ML by over-analogizing with human behavior. Anthropomorphic reasoning correctly gets a bad rap in ML, because it's very intuitively persuasive but has a mixed at best track record. This isn't a reason to abandon the human anchor, but it means we shouldn't be entirely satisfied with it.

This brings us to a third anchor, the **optimization anchor**, which I associate with the "Philosophy" or thought experiment approach that I've described previously. Here the idea is to think of ML systems as ideal optimizers and ask what a perfect optimizer would do in a given scenario. This is where Nick Bostrom's colorful description of a paperclip maximizer comes from, where an AI asked to make paperclips turns the entire planet into paperclip factories. To give some more prosaic examples:

> The optimization anchor would correctly predict imitative deception (Lin et al., 2021), since a system optimized to produce high-probability outputs has no intrinsic reason to be truthful.
>
> It also would observe that power-seeking is instrumentally useful for many different goals, and so predict that optimal policies (as well as sufficiently powerful neural networks) will tend to do so (Turner et al., 2021).

Ideas produced by the optimization anchor are often met with skepticism, because they often contradict the familiar current ML anchor, and they don't benefit from the intuitive appeal of the human anchor. But the differences from these other two anchors are precisely what make the optimization anchor valuable. If you (like me) feel that both the current ML and human anchors paint an incomplete picture, then you should want a third independent perspective.

The optimization anchor does have limitations. Since it abstracts ML into an ideal optimizer, it ignores most on-the-ground facts about neural networks. This can lead to underconstrained predictions, and to ignoring properties that I think will be necessary for successfully aligning ML systems with humans. I'll say more about this later, but some particularly important properties are that neural networks often generalize in "natural" ways, that we can introspect on network representations, and that training dynamics are smooth and continuous. Researchers focused on the optimization anchor don't entirely ignore these facts, but I think they tend to underemphasize them and are overly pessimistic as a result.

The Value of Thought Experiments

The optimization anchor points to the value of thought experiments more generally. While it poses the thought experiment of "What if AI were a perfect optimizer?", there are many other thought experiments that can provide insights that'd be hard to obtain from the ML or human anchors. In this sense thought experiments are not a single anchor but a generator for anchors, which seems pretty valuable.

One thought experiment that I particularly like is: *What happens if most of an agent's learning occurs not during gradient descent, but through in-context learning*[1]*?* This is likely to happen eventually, as ML agents are rolled out over longer time horizons (think artificial digital assistants) and as ML improves at in-context learning. Once this does happen, it seems possible that agents' behavior will be controlled less by the "extrinsic" shaping of gradient descent and more by whatever "intrinsic" drives they happen to have[2]. This also seems like a change that could happen suddenly, since gradient descent is slow while in-context learning is fast.

It would be great if we had a community of researchers making thought experiments with clearly stated assumptions, explaining in detail the consequences of those assumptions and ideally connecting it to present-day research.

Other Anchors

There are many other anchors that could be helpful for predicting future ML systems. **Non-human animal behavior** could provide a broader reference class than humans alone. **Evolution** and **the economy** are both examples of powerful, distributed optimization processes. I am most excited about better understanding **complex systems**, which include biological systems, brains, organizations, economies, and ecosystems and thus subsume most of the reference classes discussed so far. It seems to me that complex systems have received little attention relative to their germaneness to ML. Indeed, emergence is itself a concept from complex systems theory that is useful for understanding recent ML developments.

Limitations of Thought Experiments

I've focused so far on *predicting* problems that we need to address. But at some point we actually have to *solve* the problems. In this regard thought experiments are weaker, since while they often point to important big-picture issues, in my view they fare poorly at getting the details right, which is needed for engineering progress. For instance, early thought experiments considered a single AI system that was much more powerful than any other contemporary technologies, while in reality there will likely be many ML systems with a continuous distribution of capabilities. More recent thought experiments impose discrete abstractions like "goals" and "objectives" that I don’t think will cleanly map onto real ML systems. Thus while thought experiments can point to general ideas for research, even mapping these ideas to the ontology of ML systems can be a difficult task.

As a result, while we can't blindly extrapolate empirical trends, we do need a concerted empirically-based effort to address future ML risks. I'll explain why I think this is possible in a later post, but first I'll take us through an example of "taking a thought experiment seriously", and what it implies about possible failure modes of ML systems.

[1] In-context learning refers to learning that occurs during a single "rollout" of a model. The most famous example is GPT-3's ability to learn new tasks after conditioning on a small number of examples. ↩

[2] While this statement borders on anthropomorphizing, I think it is actually justified. For instance, depending on the training objective, many agents will likely have a "drive" towards information-gathering, among others. ↩

Jacob Steinhardt ▾

1 Comments

Sign in to join the conversation.

Noemi Chulo 7 months ago

I was aware of your research but not of this blog before today - this is incredibly cool. I've had many of these same thoughts before but you portray them much more eloquently than I do. Wonderful post.

♡ 0

Powered by Cove

← **Previous Post** **Next Post** →

Bounded Regret

Powered by Ghost