Frequent arguments about alignment

john-schulman

June 23, 2021

Here, I'll review some arguments that frequently come up in discussions about alignment research, involving one person skeptical of the endeavor (called Skeptic) and one person advocating to do more of it (called Advocate). I mostly endorse the views of the Advocate, but the Skeptic isn't a strawman and makes some decent points. The dialog is mostly based on conversations I've had with people who work on machine learning but don't specialize in safety and alignment.

This post has two purposes. First, I want to cache good responses to these questions, so I don't have to think about them each time the topic comes up. Second, I think it's useful for people who work on safety and alignment to be ready for the kind of pushback they'll get when pitching their work to others.

Just to introduce myself, I'm a cofounder of OpenAI and lead a team that works on developing and applying reinforcement learning methods; we're working on improving truthfulness and reasoning abilities of language models.

1. Does alignment get solved automatically as our models get smarter?

Skeptic: I think the alignment problem gets easier as our models get smarter. When we train sufficiently powerful generative models, they'll learn the difference between human smiles and human wellbeing; the difference between the truth and common misconceptions; and various concepts they'll need for aligned behavior. Given all of this internal knowledge, we just have to prompt them appropriately to get the desired behavior. For example, to get wise advice from a powerful language model, I just have to set up a conversation between myself and "a wise and benevolent AI advisor."

Advocate: The wise AI advisor you described has some basic problems, and I'll get into those shortly. But more generally, *prompting an internet-trained generative model* (like raw GPT-3) is a very poor way of getting aligned behavior, and we can easily do much better. It'll occasionally do something reasonable, but that's not nearly good enough.

Let's start with the wise AI advisor. Even if our model has internal knowledge about the truth and human wellbeing, that doesn't mean that it'll act on that knowledge the way we want. Rather, the model has been trained to imitate the training corpus, and therefore it'll repeat the misconceptions and flaws of typical authors, even if it knows that they're mistaken about something.

Another problem with prompting is that it's a an unreliable method. Coming up with the perfect prompt is hard, and it requires evaluating each candidate prompt on a dataset of possible inputs. But if we do that, we're effectively training the prompt on this dataset, so we're hardly "just prompting" the model, we're training it (poorly). A nice recent paper studied the issue quantitatively.

So there's no getting around the fact that we need a final training step to get the model to do what we want (even if this training step just involves searching over prompts). And we can do much better than prompt design at selecting and reinforcing the correct behavior.

1. Fine-tune to imitate high-quality data from trusted human experts

2. Optimize the right objective, which is usually hard to measure and optimize, and is not the logprob of the human-provided answer. (We'll need to use reinforcement learning.)
3. Leverage models' own capabilities to help humans to demonstrate correct behavior and judge the models' behavior as in (1) and (2). Proposals for how to do this include debate, IDA, and recursive reward modeling. One early instantiation of this class of ideas involves retrieving evidence to help human judges.

Honing these techniques will require a lot of thought and practice, regardless of the performance improvements we get from making our models bigger.

So far, my main point has been that just prompting isn't enough – there are better ways of doing the final alignment step that fine-tunes models for the right objective. Returning to the original question, there was the claim that alignment gets easier as the models get smarter. It does get easier in some ways, but it also gets harder in others. Smarter models will be better at gaming our reward functions in unexpected and clever ways – for example, producing the convincing illusion of being insightful or helpful, while actually being the opposite. And eventually they'll be capable of intentionally deceiving us.

2. Is alignment research distinct from capabilities research?

Skeptic: A lot of research that's called "alignment" or "safety" could've easily been motivated by the goal of making an AI-based product work well. And there's a lot of overlap between safety research, and other ML research that's not explicitly motivated by safety. For example, RL from human preferences can be used for many applications like email autocomplete; it's biggest showcase so far is summarization, which is a long-standing problem that's not safety-specific. AI alignment is about "getting models to do what we really want", but isn't the rest of AI research about that too? Is it meaningful to make a distinction between alignment and other non-safety ML research?

Advocate: That's true that it's impossible to fully disentangle alignment advances from other capabilities advances. For example, fine-tuning GPT-3 to answer questions or follow instructions is a case study of alignment, but it's also useful for many commercial applications.

While alignment research is useful for building products, it's usually not the lowest-hanging fruit. That's especially true for the hard alignment problems, like aligning superhuman models. To ensure that we keep making progress on these problems, it's important to have a research community (like Alignment Forum) that values this kind of work. And it's useful for organizations like OpenAI to have alignment-focused teams that can champion this work even when it doesn't provide the best near-term ROI.

While alignment and capabilities aren't distinct, they correspond to different directions that we can push the frontier of AI. Alignment advances make it easier to optimize hard-to-measure objectives like being helpful or truthful. Capabilities advances also sometimes make our models more helpful and more accurate, but they also make the models more potentially dangerous. For example, if someone fine-tunes a powerful model to maximize an easy-to-measure objective like ad click-through rate, or maximize the time users spend talking to a chatbot, or persuade people to support a political agenda, it can do a lot more damage than a weak model.

3. Is it worth doing alignment research now?

Skeptic: It seems like alignment advances are bottlenecked by model power. We can't make useful progress on alignment until our models are powerful enough to exhibit the relevant phenomena. For example, one of the big questions is how to align super-human models that are hard to supervise. We can think about this problem now, but it's going to be pretty speculative. We might as well just wait until we have such models.

Advocate: It's true that we can make faster progress on alignment problems when we

can observe the relevant phenomena. But we can also make progress preemptively, with a little effort and cleverness. What if we don't?

- In the short term, companies will deploy products that optimize simple objectives like revenue and engagement. Doing this responsibly while looking out for customer wellbeing is harder and requires a more sophisticated methodology, including alignment techniques. Unless the methodology is well-known, companies won't bother, and no public or regulatory pressure can force them to use something that doesn't exist.
- In the long term, aligning superhuman AIs might end up being very hard, and it might require major conceptual advances. If those advances aren't ready in time, the consequences could be severe. If AI progress continues at its current fast pace, we might end up in a situation that AI is so useful, that we're obligated to use it to run companies and make many decisions for us. But if we don't have sufficiently aligned AI by that point, then this could turn out badly.

In my experience, the "tech tree" of alignment research isn't bottlenecked by the scale of training. For example, I think it's possible to do a lot of useful research on improving the RL from human preferences methodology using existing models.

Is there an even better critique that the Skeptic could make? Are Advocate's responses convincing? Does the Advocate have a stronger response to any of these questions? Let me know.

Thanks to Avery Pan, Beth Barnes, Jacob Hilton, and Jan Leike for feedback on earlier drafts.