Your posts should be on arXiv

JanBrauner

August 25, 2022

TL;DR: There are many posts on the Alignment Forum/LessWrong that could easily be on arXiv. Putting them on arXiv has several large benefits and (sometimes) very low costs.

Benefits of having posts on arXiv

There are several large benefits of putting posts on arXiv:

- 1. Much better searchability, shows up in google scholar searches.
- 2. Additional reads (arXiv sanity, arXiv newsletters, and so on).
- 3. The article can accumulate citations, which are shown in google/google scholar search results.

1) • 3) lead to more people reading your research, which hopefully leads to more people building on it and maybe useful feedback from outside of the established alignment community. In particular, if people see that the paper already has citations, this will lead to more people reading it, which will lead to more citations, and so on.

You'd gain even more of 2) and 3) from publishing it at a conference, but, unlike arXiv, that's significant additional work (often still worth it).

There are also some smaller benefits from publishing on arXiv:

- 4. firmly establishes this as your contribution (not sure, but I think if you only have an alignment forum post, someone could build a bit on it and then claim the whole thing as their contribution because alignment forum posts don't count?).
- 5. better citability (e.g. if somebody writes an ML paper to be published in ML venues, it gives more credibility to cite arXiv papers than Alignment Forum/LessWrong posts. The same goes for people writing e.g. Wikipedia articles about alignment.)

How much work is it to submit to arXiv?

Sometimes, I think it can be as easy as creating a pdf of your post and submitting it (although if your post was written in LaTeX, they'll want the tex file). If everything goes well, this takes less than an hour.

However, if your post doesn't look like a research article, you might have to format it more like one (see this comment). Ultimately, I'm not 100% sure which content format exactly arXiv does and doesn't accept; so it'd be good for people to report their experiences in the comments.

If you're a Very Busy Alignment Researcher, I'm sure you can outsource large parts of this. E.g. FAR's comms staff could probably help. I also have worked with a freelancer on similar things in the past (like making publications look nice in LaTeX), feel free to reach out for the contact data.

What types of posts should be on arXiv?

I'd say, submit anything that fits on arXiv/wouldn't look out of place there. This probably includes most research contributions or otherwise research-y or academic-y things. Submitting to arXiv is particularly useful if the post's target audience is wider than the LessWrong readership.

Here are some examples from posts that, IMO, clearly should be on arXiv (skewed by what I read and remembered):

- Joe Carlsmith's Draft report on existential risk from power-seeking AI (it's on arXiv now, but >1 year after it was published)
- Neel Nanda's and Tom Lieberum's A Mechanistic Interpretability Analysis of Grokking (Neel has submitted it to arXiv after I hassled him to do it :-P)
- High-stakes alignment via adversarial training [Redwood Research report] (was on arXiv to begin with, good!)
- I haven't read it, but from the name of it, it sounds as if ARC's first technical report: Eliciting Latent Knowledge should be on arXiv
- Ajeya Cotra's Draft report on AI timelines
- Richard Ngo's AGI safety from first principles
- The Truthful AI work by Lin, Evans, Cotton-Barratt, and others (several papers, all on arXiv)
- …

If arXiv doesn't fit

There are also some other posts on the Alignment Forum/LessWrong whose target audience is the wider AI community. I think should be published additionally elsewhere. A great example is (one of my all-time favourite posts) Ajeya Cotra's Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. This maybe wouldn't really fit on arXiv (although it wouldn't be crazy to put on on arXiv either). But there is a range of other venues that might publish it, from Towards Data Science (on the low effort, low prestige end) to the MIT Technology Review (on the high effort, high prestige end).