How I am transitioning from Physics to AI safety research

Oz Amram

Sep 22nd, 2026

Hello! If you are reading this I am assuming you are interested in getting into AI safety research, perhaps coming from a physics background. I am currently navigating this transition, I don't have a position lined up, and I definitely don't have all the answers. But I have been researching / reading a lot the last few months to prepare and I am happy to share what I found.

Resource: LLMs

A piece of meta advice I would give is use LLMs to help you! LLMs are extremely useful for learning about new things, and navigating this transition I have been making use of them frequently! At least a $20/mo subscription is absolutely worth it. Ask it for a reading list of papers in some area, have it load a paper into context and you ask questions about it, have it ask *you* questions about a paper to test your knowledge. Have it search for fellowships / opportunities for you. For writing applications, check the specific policy of whatever you are applying to and follow that, but my general thought these days is DON'T have an LLM write anything for you, but DO write something yourself, have an LLM critique you / give comments, and then you integrate the feedback you like into what you wrote. They can be a bit pedantic, and aren't always right but are extremely useful at clarifying your thinking / writing or pointing out if you missed something major.

Resource: 80,000 hours (website)

80,000 hours is a non-profit organization to help people transition into high-impact careers that help the world. One of their focus areas is AI safety (link).
They have a good overview of the risks, and resources to help people make the transition, including a job board for AI safety roles, an AI chatbot designed to help discuss this type of career transition that I actually thought was pretty helpful, and you can apply for a free 1:1 career coaching session which I found very helpful. I would definitely recommend it!

Resource: Funding for Career Transitions into AI Safety

Several organizations provide funding to individuals looking to transition their career into AI safety. They can fund you to do a research project on AI safety on your own (salary and some compute) as a way to get into the field. Turn around times seem quick.

See: Transformative AI Fund, Blue Dot, and Manifund

Step 1: Get familiar with the technical AI safety literature.

This includes learning about the type of risks people worry about (and the jargon):
loss of control, AI-enabled biological risks / pandemics, threat models, takeover, cyber risk, scheming, reward hacking
Make sure you understand the basics of LLMs: transformers, scaling laws, pretraining vs posttraining, RLHF, RLVR, chain of thought, etc

Learn the broad strokes about what the different types of technical safety work:

- Threat modeling
- Evaluations of alignment, multi-agent behavior, scheming, etc
- Current practice alignment methods, and their failures
- Mathematical / more theoretical approaches to alignment (eg ARC and similar)
- Mechanistic interpretability
- AI policy and governance

ARENA has a full curriculum you can check out (link)

If you want to dive in, LessWrong and the Alignment Forum, where much of the field posts its technical work, and 'takes'.
Twitter is also super helpful to learn the vibes of the community. You can learn a lot if you follow the right people (eg Roon). There are a lot of really good insights, it keeps you up with an extremely fast changing field, and it helps see breaking developments and how the community is reacting (eg new model releases, security incidents) but it will also **melt your brain** so be careful.

Once you roughly know the broad strokes of the different areas, pick one that looks interesting to you and read more in depth. Ask an LLM for a reading list on some topic and start reading papers. Follow the thread of whatever is interesting to you. Because it has developed so quickly, AI safety is not as deep as a topic like physics (you can read ~5 papers on a topic and basically understand what's going on), but it is very broad (there are many many topics) and things are moving very fast. I think it rewards people who are self-directed learners, which if you have done a PhD you should be. This stage is about learning and seeing what topic you might be well suited for.

Step 2a: Apply to things

Once you have a basic grasp of what type of research you might want to do, start applying to things.

Fellowships

Your best bet is probably fellowships meant to transition people from other technical fields into AI safety. These programs do not require prior experience in AI safety research, they are trying to get people from fields like math / physics / ML / statistics into AI safety.
I applied to MATS, CBAI, LASR (deadlines for fall applications have passed, but I expect there to be another round in a few months)
There is also: Anthropic Fellows, Astra, MARS, Pivotal, SPAR, PIBBSS, a curated fellowship list here
Ask an LLM to help you find more. Note that these vary in location, (London or Bay Area are the hubs), a few are remote + part time, most but not all are full time, they might differ in the topics covered by their mentors. Apply to what works for you

Labs

If you have some relevant ML background you can also apply directly to AI labs
Eg Anthropic has open roles for research scientists and engineers in Alignment and Interp at the moment.

In addition to the Frontier labs, there are also a bunch of 3rd party research orgs: METR, Redwood, Apollo, Transluce, ARC, Resolution, UK AISI, RESI, Simplex, …
I believe these 3rd party labs can definitely have an impact! Both in creating an ecosystem of independent monitors to do evaluations (which we definitely need). The 3rd party orgs vary in their focus, some are deep into theoretical aspects of alignment, some are doing evaluations of current models, some are doing empirical alignment or mech interp research

Check out this 'map' of the AI safety ecosystem! https://aisafety.com/map
There are orgs more focused on 'theoretical research', empirical research, governance, etc

Research from these 3rd party labs definitely affects what the frontier does
Eg this paper: https://arxiv.org/abs/2502.17424 and its follow ups, was led by 3rd party people, and had huge impact
Eg Anthropic studying the same effect at scale:
https://www.anthropic.com/research/emergent-misalignment-reward-hacking
And presumably adopting some of the mitigations into their alignment practices

Step 2b: Start getting your hands dirty

I think it's important to start doing things asap, research can only be learned in practice. This is also step two because it should be done in parallel with your applications I think. No need to wait.

You can start working through material from the ARENA curriculum or apply to go to one of their (funded) bootcamps!

Brainstorm an idea of some project based on one of the papers you read, perhaps something small to start, and just try things out. AI agents definitely lower the technical barrier to entry to do things. Post your idea / progress on forum like LessWrong to get feedback as soon as you can. You can also cold email researchers who are working in that area to get their thoughts, people are friendly. Having some concrete project to talk about in your applications helps, and it starts the learning process.

One excellent AI safety researcher Neel Nanda has a very nice outline of what he thinks are interesting research topics that you can use to get started with idea generation (link, click Recommended Research Topics on the right sidebar)

Once you have something worth sharing, even if not a full paper, you can post it to LessWrong (a central forum for AI safety topics, many top researchers frequent it) to get feedback! And many applications in this space take a well-crafted LessWrong post seriously as a research product. Make the code public on github too.

Outlook

There is a big influx of philanthropic funding into AI safety at the moment (eg Coefficient Giving is directing ~$1B in 2026). So I expect many existing orgs to be hiring and new orgs to spring up soon!

Currently there are probably ~2k AI safety researchers working right now in the world (as compared to ~20k working particle physicists) so it definitely seems possible to have an impact!