General advice

There isn't a standard career path in this area. AI alignment is a pre-paradigmatic field in which nobody has a good idea what the right prerequisite knowledge is or what an answer looks like. That means this is a path for people who are willing to wrestle with uncertainty.

Financially supporting your research can be hard; funding isn't a reliably solved problem, but opportunities for funding do exist.

Rather than thinking of your goal as trying to "become a researcher", it might be better to think of it as trying to solve the alignment problem. You can get started by reading and thinking about the problem, maybe commenting on posts, and writing down your own ideas in private docs or on LessWrong. Don't necessarily rely on getting feedback without actively reaching out to people who might have good thoughts. It will help you to find peers to be in contact with.

One way to get into conceptual work is by writing distillations of other people's work, or critiquing key posts in places like LessWrong (which includes everything that has been posted on the Alignment Forum). It's important to develop your own "inside view" on the problem.

Consider asking around your personal network for an alignment research mentor, or a collaborator who knows the literature and can give you pointers and feedback. This is unlikely to work with leading alignment researchers, who already get a lot of requests for mentorship, but may be more likely to succeed with people you locally know who can teach you generic research skills. It depends a lot on the person. If you can get a mentor, that's great, but you don't need one to succeed, so don't get blocked on it: almost everything you can get from a mentor, you can also get from a mix of learning by doing and having discussions with and getting feedback from peers. It will take you a bit longer and you'll probably hit a few more dead ends without a mentor to guide you, but you can do it.

Training programs

Consider training programs (e.g. SERI-MATS) and internships. AI Safety Training has an overview of these. AGI Safety Fundamentals runs courses on AI alignment and governance. The 80,000 Hours AI safety syllabus lists a lot of reading material. For more suggestions, look at Linda Linsefors's collection of do-it-yourself training programs.

If you're applying to a program, choose whichever one you think you will most enjoy. The important thing is to start learning the field and to get some contacts. You'll end up learning different things in different programs, but you won't be locked into that path. You're free to continue exploring whatever direction you want and to apply to other programs in the future, and you'll have a much easier time navigating the space when you have some context and some connections.

Guides and resources

Some helpful guides:

- MIRI has a field guide for doing alignment research, especially in the context of a research group like MIRIx.
- How to pursue a career in technical AI alignment by CharlieRS has a section with advice on how to get into theoretical AI alignment work.
- Rohin Shah has an FAQ on AI alignment career advice.
- John Wentworth has a post on getting into independent alignment research.
- There is also Adam Gleave's Careers in Beneficial AI Research document.
- Richard Ngo has posted on AGI safety career advice.

Other resources:

- 80,000 Hours offers calls where they give career advice. AI Safety Quest also provides advice in navigation calls.
- AI Safety Support has a lot of other good resources: their links page, Slack, and newsletter.
- See AI Safety Ideas for research ideas.

Related

- I'd like to be a lead researcher / mentor / advisor. How can I do this?
- What are some AI alignment research agendas currently being pursued?
- How can I use a background in the social sciences to help with AI alignment?

Scratchpad

We're still working on this answer. Check back later, and in the meantime, see AI Safety Support for advice.

From old doc:

Write distillations

Write objections to key posts

Form an inside view

John Wentworth getting into independent alignment research

JJ how to get started in AI alignment

AISS

80k coaching

Consider doing a PhD

Go to EAG

Training programs / internships / etc

Ask around for an alignment mentor

Doing distillations is good

Build an inside view

SERI MATS program

AISCamp

Training programs

AGISF

aisafety.ideas

John Wenstworth how to get into independent alignment research

Rohin Shah's posts

Consider doing a PhD - link

AISS

For a comprehensive overview of AI safety programs and organizations, see http://aisafety.world/

Also consider: I'd like to become an AI alignment researcher. How should I go about this? - Response there (state Jan 2023):

AI Safety Support offers free calls to advise people interested in a career in AI Safety, so that's a great place to start. We're working on creating a bunch of detailed information for Stampy to use, but in the meantime check out these resources:

- EA Cambridge AGI Safety Fundamentals curriculum
- 80,000 Hours AI safety syllabus
- Adam Gleave's Careers in Beneficial AI Research document
- Rohin Shah's FAQ on career advice for AI alignment researchers
- AI Safety Support has lots of other good resources, such as their links page, slack, newsletter, and events calendar.
- Safety-aligned research training programs (under construction).

Linda's thoughts:

- Combine applied and conceptual work, because the advice is pretty much the same.
- List various training they can apply to, and pros and cons with each. Including how competitive they are and what type of background is expected, etc.
    - AGISF
        -
    - ML Safety
    - AISC
    - SERI MATS
    - PIBBSS
    - etc…
    - For the most recent list of programs and for application dates, go to AI Safety Training for dates, applications, most recent take
- Advise: "Apply to whichever program you think you will enjoy. The important thing is to start learning the field and to get some contacts. You will end up learning different things in different programs, but you will not end up being locked into that path. You are free to continue exploring whatever direction you want and apply to other programs in the future, and you will have a lot easier time navigating the space when you have some context and some connections.
- Other options
    - "I don't yet have the background required for any of those programs." → Go to [?]
        - "That's OK. We all come from different backgrounds. Book a call with AISS to figure out what's the best way for you to contribute.
        - (The people who end up here will be few enough and diverse enough such that it's better they just talk to a person)
    - "I prefer to self-study (first)" → Go to [?]
        - AGISF curriculum
        - AF sequences
        - etc…
    - "Those programs look good but I did not get in or the timing doesn't work for me" → Go to [?]
        - [Draft] Do-it-yourself versions of some AI Safety Programs - Google Docs

- “Is there some shorter event I can attend?” → Go to [?]
  - EAG and EAGx and LW events are great places to meet others interested in AI Safety
  - Some academic conferences have AI Safety workshops
  - Alignment Jam
  - Other events: AI Safety Training

FQ: I’d like to be a lead researcher / mentor / advisor. How can I do this? **[2.0.2]**

FQ: Read AF / AGISF / Read Alignment Newsletter