Part 1: Excerpt from Gleave's Careers in Beneficial AI Research

Is Technical AI Research Right for You?

This document primarily focuses on careers in technical AI research, since this is the area I am most familiar with. I believe this can be one of the highest impact careers for people with a technical background. For example, 80,000 Hours has a dated but still useful summary of one reason to work in this area (reducing long-term risks). [Note: an updated version can be found here.]

However, there are many other promising areas to work on. Ensuring beneficial AI is developed also requires work in AI strategy and policy. While this may seem less of a natural path for those with a CS background, note that strategic work is also enriched by having some people with a technical background.

Skills within Technical AI Research

If working on technical AI research is the right fit for you, then there are four main relevant skill sets:

A. Software engineering: infrastructure, building environments, etc.
B. ML implementation: converting a research idea into a working model.
C. ML research direction: coming up with good ideas, designing experiments.
D. Theory research: building good abstractions, mathematical reasoning.

Any mapping from job titles to skills is necessarily approximate, but in general:

- Research engineers have a lot of B and some of A.
- Deep learning (e.g. ICLR, much of NeurIPS) researchers need to be competent at both B and C; the weighting varies.
- Other subfields of ML (e.g. most of the non deep learning NeurIPS subject areas) tend to be mostly C with some D.
- Theory, such as computational learning theory (e.g. COLT) or MIRI's Agent Foundations agenda, is primarily D.

Those with a strong background in A can typically learn B, but it is a very different style of development which takes time to get used to (now-DeepMind engineer Matthew Rahtz' experience is representative). Those with prior experience in numerical programming, an aptitude for applied mathematics and who are used to working with messy codebases are likely to find the transition easier.

Learning C or D tends to take a lot of time, and involves working with experienced researchers either in a PhD program or an industrial lab. It will be easier for people with prior research experience in STEM subjects.

Progress in beneficial AI is bottlenecked on both experienced researchers and research engineers. Most groups I've spoken to in industry have a slight preference for additional researchers, but this is highly sensitive to personal fit: they'd rather hire a great research engineer than an average researcher.

Part 2: Deeper dives

Read whichever 2 or 3 of the following seem most relevant to you. Broadly speaking, people with less programming or ML experience should read the earlier ones; people with more research experience should read the later ones.

1. How to start coding
2. AI safety needs great engineers (Jones, 2021): what skills are sought-after by large engineering companies working on safety?
3. ML engineering for AI safety and robustness (Olsson, 2018): as above
4. Guide to working in AI policy and strategy (80,000 hours, 2017)
5. How I formed my own inside views about AI safety (Nanda, 2022): advice for coming up with good models to direct research ideas that solve real problems
6. How to pursue a career in technical AI alignment (Rogers-Smith, 2022)
7. PhD application advice (Gleave, 2020)
8. A survival guide to a PhD (Karpathy, 2016)
9. A recipe for training neural networks (Karpathy, 2019)
10. Research as a stochastic decision process (Steinhardt, 2019): advice for doing excellent research
11. Research taste exercises (Olah, 2021): as above

Part 3: Your own career

Spend 5-10 mins (set a timer!) brainstorming how the ideas from these readings are relevant to your own career. Maybe that involves actually doing the research taste exercises, or thinking about the next steps to take to apply to jobs/PhDs.

When in doubt, prioritize upskilling - compared with other fields, ML and alignment are both unusually meritocratic, and skilled people can do very well even without credentials. Prioritizing upskilling may involve having less direct impact in the early years of your career. In general this is fine, since work done by more skilled researchers is much more valuable, as long as you feel confident you won't drift away from your current intentions.

If you feel you want to go into more depth, you might undertake the 80k career planning course (potentially as a capstone project). If you want to talk through your subsequent plans with expert careers advisors, 80,000 Hours would be delighted to see your application for a careers advice call.

Some default actions that you might plan to do (possibly as part of the capstone project):

- Take a programming course
    - Start with Python - e.g. https://www.learnpython.org/
- Apply for coding bootcamps
- Take a ML course
    - Default option: Fast.ai's Practical Deep Learning for Coders
    - Others:
        - Google's machine learning crash course
        - Stanford computer vision course (2017)
        - NYU deep learning course
        - Spinning up in deep RL
        - Hilton's Deep Learning curriculum
- Replicate some ML papers
    - See this list of key deep RL papers (although if you haven't implemented neural networks before, start with supervised learning instead)
- Start writing about alignment (e.g. writing informal reviews of papers, etc)
- Apply for research internships
    - Most will be available in mainstream ML, so those should be your main focus. A couple more specifically focused on alignment:
        - CHAI internships
        - DeepMind internships
- Apply for AI residencies
- Apply for PhDs. Some groups interested in taking students for alignment work:
    - CHAI
    - David Krueger at Cambridge
    - Sam Bowman at NYU
    - Jacob Steinhardt at Berkeley
    - Tim Oates at UMBC
    - Roger Grosse at UToronto
    - Dylan Hadfield-Menell at MIT
    - Professors listed in Future of Life Institute webpage
- Apply for funding if that would help with any of these things
    - Open Philanthropy AI PhD scholarship
    - Open Philanthropy early career funding
    - Open Philanthropy undergraduate scholarship
    - Long-term future fund
    - Survival and flourishing fund

- The Center on Long-Term Risk Fund (CLR Fund)
- Future of Life grants