Careers in Beneficial AI Research

Adam Gleave

Last updated: 2020-07-06

Introduction

This document is intended for people considering pursuing a career in AI research and who want to ensure AI has a beneficial impact on society. I start by considering an overview of possible careers in this space and briefly discuss the trade-offs. I then discuss how to test your personal fit for different careers, and how to best demonstrate your abilities. The final part of the document focuses on tips for PhD applications.

I'm a PhD candidate at UC Berkeley, and have been involved with PhD admissions and interviewing prospective research engineers. This document seeks to address the questions I've most frequently been asked by prospective researchers.

This document is meant to express **only my personal opinion**. I am grateful for feedback from Rohin Shah, Dan Hendrycks, and several anonymous reviewers -- but naturally, any errors are my own. I would encourage you to consult some of the other excellent sources of advice online (many of which are linked to in this document) to obtain a broader perspective.

Choosing A Role: The Beneficial AI Landscape

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).

However, there are many other promising areas to work. 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.

If your motivation for AI is to ensure the long-term flourishing of humanity, it is also worth considering how to reduce biorisk, extreme climate change or identifying currently unrecognized problems via methods such as horizon scanning.

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 (Matthew Rahtz experience is representative). Those with prior experience in numerical programming, an aptitude for applied maths 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.

Preparing for a Role

Research Engineering

The best way to learn research engineering is to work somewhere there is both high-quality engineering and cutting-edge research. Apply to residency programs at industrial labs. The top-4 labs are DeepMind, OpenAI, Google Brain and Facebook AI Research (FAIR); there are also smaller (but good) safety-focused labs Anthropic and Redwood Research. There are also many smaller players like Amazon AI, NVidia, Vicarious, etc. These are generally less desirable, but still good options.

Any of these labs will have some amazing teams, so the best choice for you depends to a large extent what kind of agenda you want exposure to. For example, OpenAI's research often involves scaling up deep learning techniques with large amounts of compute and data, and they have operated across a range of domains (RL, NLP, CV, etc). By contrast, much of FAIR's research focuses on NLP, but they employ a broader range of techniques. With that said, don't worry too much about whether a group's agenda fits your personal interests, or even if you think it's useful -- once you have some experience, it'll be easy to move between teams and labs.

Many of these labs hire software engineers to work on infrastructure and tooling in addition to research engineers. These roles will tend to accelerate research progress across the company as a whole. That's good *when* the company is focused on beneficial AI research, is likely close to neutral for more application oriented companies, and may be net-negative if the company is working towards AGI but does not have a strong safety commitment. It's often possible to transition over time into a more RE-style role, although this varies by company (my weak impression is it's likely to be easier at smaller places).

To make yourself a more attractive candidate, you can build up your skills by reproducing papers. Blogging and open-sourcing your code is a great way to build up a portfolio.

It's also possible to build up these skills in academia, whether via research internships or a PhD. This route is far from ideal, however, as your advisor will be primarily concerned with research direction and your peers will be at best adequate engineers.

Research

I would encourage most people interested in research to pursue a PhD. It is also possible to pursue a research career without a PhD, e.g. via a residency program at an industrial lab, or starting work as a research engineer and gradually transition to a more research-oriented role. This may be faster than completing a PhD, although may still involve as much (and, in the case of research engineering, almost certainly more) non-research activities as a PhD. Moreover, it will typically limit your options, as many organisations still insist on a PhD. OpenAI and Google Brain are amongst the most flexible industrial labs, and MIRI also hires people without PhDs. I would encourage everyone to apply to a residency program nonetheless -- you'll learn a lot and can always start a PhD a year later.

How Do I Get Research Experience?

If you're fortunate enough to be an undergraduate at a top research institution, then take advantage of it and reach out to people at your university. Often you'll find PhD students and post-docs more responsive than professors. However, make sure you get a professor to agree to take you on officially, and try to meet with them regularly: you'll want to get a recommendation letter from them when it is time to apply.

If you haven't already done a Master's, this is one of the simplest ways of gaining research experience. Make sure the program focuses on research; many Master's degrees, especially in the US, are focused more on preparation for industry. Note that the bar for admission will go up post-Masters, with PhD programs expecting you to have significant research experience including some publications.

An equally good, if not better option, is to intern in a machine learning group. Getting these positions is typically more competitive than Master's degrees, however. There are a handful of programs offering short-term research experience, notably:

- CMU RISS -- only available for undergraduates.
- MILA.
- TTIC Visiting Students: most faculty participating in the program have an interest in ML.
- ICL UROP: established program for undergraduate research opportunities. Some advertised opportunities; can also approach individual lecturers (see their ML group).
- Aalto University Science Institute has an internship program, with some ML projects
- LLNL Data Science Summer Institute -- offers some (applied) ML projects.

Most advertised internship opportunities are over the summer, but in practice groups are often flexible regarding when they take interns, so it's worth enquiring if the advertised dates don't line up with your availability.

There are also many groups that will be open to taking interns or hiring for one to two year research assistant roles. However, professors are inundated with generic cold e-mails. Ideally try and get introduced, for example by a professor who taught you or you used to work with. Failing that, your best bet is to understand their research in detail and send a targeted message. Again, it may be better to start with post-docs or PhD students.

Some labs focusing on beneficial AI also offer internships:

- CHAI
- FHI

The main advantage of these programs is you will develop a deeper understanding of what beneficial AI research looks like. If you impress it will also of course help you get into that lab. However, it may be worse preparation for applications than other internships: since beneficial AI is a new and rapidly growing field, mentorship is often more limited, and it can sometimes be harder to publish on the topics.

There is also the AI Safety Camp, who put together teams to work on a safety-related project proposed by a mentor. I think this is a good option if you're already in a PhD program and looking to shift your agenda more towards a safety focus, as the project ideas are often good, and you'll be working with similarly motivated people. However, my impression is that the mentorship is generally fairly hands-off relative to a typical internship, so requires more independence. Additionally, the projects may be harder to publish (or at least take more time to publish), and the mentors are often not in a good position to write recommendation letters, which is problematic if you are applying to PhDs.

It is important to also put time into impressing your adviser: after publications, the most important aspect of your application are recommendation letters. Although the best way to do this is to do great research, preparing carefully for your meetings will also help.

PhD Application Advice

The Process

Applying to grad school to work on beneficial AI is, for the most part, like applying to grad school in any other area of CS.. I would therefore recommend first reading these general resources:

- Applying to PhD programs in CS - Prof Harchol-Balter, CMU
- Advice on applying to CS Grad School - Prof Justine Sherry, CMU

Note that both of these describe the application process for US PhD programs in CS. Conventions are different around the world. In particular, in the UK one is expected to make

contact with a faculty member prior to applying, and work out a research topic. One then writes a fairly detailed (2-3 page) research proposal on that topic in the application.

If you don't have time to read the two documents, here's the most important thing to know: **what determines whether you get into a top program is your research experience**. This is measured by your publications and recommendation letters. You are extremely unlikely to get into a top 5 program unless you have at least one publication. You should satisfice on your grades: make sure they are OK, and put the rest of your time into research. Stanford has a minimum GPA requirement of 3.6, and is among the more stringent of programs. Having very strong grades (close to 4.0 GPA and/or ranking high in your year) will help somewhat, but it is not the best use of your time. Note grades in relevant courses are slightly more important as some reviewers will skim your transcript.

In much of this document, we will discuss ways to hone your statement of purpose and other aspects of your application to maximise your chances, but remember: by the time you write your application, your odds of getting in are largely predetermined. A very bad statement might tank an otherwise good application, and an excellent statement might drag a marginal application above the bar, but your application will largely live or die based on your research track record.

Where Should I Apply for PhDs?

Summary: aim for top groups; give more weight to program quality than the ability to pursue topics you consider important.

I'd encourage you to work with a top group, to the extent that it is likely worth taking one or two years out to gain further research experience before doing a PhD if you do not get an offer from a top group on your first attempt. This is for a few reasons:

- Only the top universities have professors covering the whole range of AI research topics. Although you can often find a great professor in a particular area at second-tier universities, this in effect means pre-committing to a research topic. It's very common for people to change topics after interacting more with the field, so the loss of option value is a significant cost.
- A lot of the value of a PhD comes from conversations with other students in the program, so the quality of your cohort is key.
- In 'hot' areas, it's hard to keep up purely by looking at published material, so you *need* to be in a top group to compete.
- Prestige seems to still matter for faculty hiring.

Note most industrial labs also offer 'residency' programs open to graduates without a PhD. I'd encourage everyone to apply, as they're a great way to quickly gain valuable skills, and can lead to a permanent job offer. They select slightly differently to PhD programs, so are a particularly good option if you have an unusual background (e.g. strong work experience, or have a PhD in another field). Programs:

- Google Brain Residency
- Facebook AI Research Residency
- Microsoft Research Residency
- OpenAI Machine Learning Fellow

Another factor is the ability to which you will be able to work on topics related to beneficial AI during your PhD. Opinions differ in how important this factor is. PhDs are a substantial time investment (min 3 years for UK, up to 6 common in the US), so there's scope to have a substantial positive direct impact from research during your PhD. However, you are likely to be much more productive as a researcher towards the end of your PhD than at the start, so it may be better to prioritize learning.

A contentious issue is the degree to which most ML research experience prepares you for research in beneficial AI. If you expect transformative artificial intelligence to be developed in the near future, then it is likely to be closely connected to current techniques, so becoming an expert in state-of-the-art machine learning seems very useful (see Paul Christiano's prosaic AI alignment argument). By contrast, if you expect most of AI's impact to lie further in the future, then it is more important to become generally competent as a researcher. Regardless of your timelines, it may also be valuable to explore beneficial AI's many interdisciplinary connections with cognitive science, social choice theory, decision theory, and other fields.

I'd recommend everyone at least keeps up-to-date on beneficial AI research during your PhD, and would strongly encourage working on it at least part-time. Most advisors will allow this amount of flexibility, especially later on in your PhD, but do scope this out ahead of time.

Fortunately, it is increasingly possible to work with great advisors on ML work that is also connected to beneficial AI. Some of the great places to work are Berkeley (e.g. CHAI, RAIL), Stanford (Sadigh, Finn) and Toronto (Grosse, Duvenaud). Oxford is also a good option due to the presence of FHI (although FHI cannot directly take students).

Even if there's no one at your university interested in beneficial AI research, it is often possible to collaborate with researchers elsewhere. Industry internships are one natural route (and often lead to collaborations that continue after the internship). It is also usually possible and often desirable to visit other academic labs.

Your Application's Audience

Typically your application will first be screened by AI graduate students. Shortlisted applications will make their way to professors who you mentioned in your application and/or are working on relevant topics. The reviewers will be skimming your application: a colleague at Berkeley reviewed 120 applications this year, putting an average of around 8 minutes per application (which are around 20-30 pages long), which he suspects is *more time* than the average reviewer will invest. Marginal applications will receive somewhat more attention than this.

You cannot control what order your application will be read in, although starting with a CV or recommendation letters is common. Important information (e.g. a publication or prestigious award) should be prominent and placed in your CV, statement of purpose *and* recommendation letters! Your goal is to pique the reviewer's interest to actually read your full application.

How to Write an Academic CV?

Be aware an academic CV is different in style to a business CV. The main difference is that academic CVs can grow very long: one or two pages would be typical for a grad school applicant, 3-5 pages for a newly minted PhD, with professors usually hitting double digits.

Emphasize your research experience and publications. You don't need to list the classes you took, unless there's something you want to draw attention to (e.g. you came top of your AI class), you will have to submit your transcript anyway.

Berkeley has a good guide to writing an academic CV. It's aimed at people who have just finished a PhD, so some of the advice is off base, but it's easy to adapt.

How Should I Write a Statement of Purpose? (US)

Also called a "personal statement" or "research statement". Section 3.3 of the CMU guide has good general advice on this, read it first.

A typical statement of purpose would start by introducing your research experience, explain why you want to do a PhD, outline your research interests and then connect this with the group you are applying to by highlighting relevant work they have conducted. Note that it is *not* a research proposal, as is needed in the UK and many other countries. See exemplars below for successful statements of purpose.

There's a few areas people often struggle with.

Formatting

You are expected to cite relevant work in your statement of purpose. Please make sure to format the citations and references correctly and consistently! This is easiest if you use LaTeX. If you don't already know how to use LaTeX, learn it if you have time; you will use it throughout your research career. Although this may be a minor point, a reviewer cannot help but be sceptical as to how prepared you are for graduate school if you do not conform to standard technical writing styles.

How Specific to be in Research Interests?

This is a trade-off. You want to demonstrate mastery of the area and the ability to identify novel promising research questions. However, if you veer too close to a research proposal, you may alienate reviewers who are not interested in your pet topic. Generally, I'd suggest first discussing interests in high-level areas (e.g. deep reinforcement learning) and then outlining specific topics. For example (do not copy this, the proposed research topic has already been done):

> "I think that the next transformative capabilities in AI will come from deep reinforcement learning, since it has seen such success in general tasks such as learning Atari games from pixels. The work in this field is new and novel and I think there is lots of opportunity for improvement, for example if instead of selecting trajectories for experience replay using a uniform distribution, we prioritize the ones which we are worst at predicting. I'm also interested in developing meta-learning and hierarchical RL approaches and applying them to these tasks."

If you are confident your reviewers will be interested in your ideas (e.g. you used to work in the group or have run your idea past someone there), then you can lean more towards a detailed research proposal. If you feel you are a marginal applicant, it may also be worth being more specific, since it will increase your variance which might push you above the bar on some applications.

How to Discuss Beneficial AI?

If you are reading this guide, you likely want to pursue research to ensure AI has a beneficial impact on society. It's certainly worth discussing briefly (2-3 sentences) why the problems you want to work on have a social impact, and how this motivates you, but it's best to avoid focusing on this too much. Remember a range of people may read your application: you do not want to inadvertently insult anyone by implying (often incorrectly) that their research field is not a "high priority" area!

Moreover, it is important to stick to technical topics, which show your ability as an AI researcher and give the reader a better sense of what you would be like as a colleague. So discussing your interest in adversarial examples or safe exploration is good; but digressions into topics such as moral philosophy are best avoided (unless applying to an interdisciplinary program). Where possible, relate your interests to areas the reader will be familiar with: e.g. concrete problems in existing systems. Also be sure to cite existing relevant work in the field: someone familiar with the field will expect it, and those unfamiliar will be more inclined to believe it's important if others have studied it.

As a rule of thumb: imagine both Yann LeCun and Stuart Russell will be reading your application, who it is fair to say have some disagreements. Try and write something that would appeal to both. For example, you can cite sources that appeal to a broad audience like Concrete Problems in AI Safety yet are still reflective of your research interest.

Example Application Materials

I am making available my application materials and those of others who have given me permission. They will be most useful to get a sense of possible styles, and inspiration for some common themes (e.g. how to discuss an interest in AI safety). Please be original and come up with your own research proposals, it is obvious when the ideas are not your own. (If you have been admitted into grad school and are willing to share your material, please e-mail me.)
Caveat: it is unclear whether we were admitted because of or in spite of these, so please exercise your own judgement.

- Adam Gleave: admitted for AI into Berkeley, UW, Oxford, Cambridge; waitlisted MIT; rejected from CMU. CV, statements for Berkeley, CMU and research proposal for Oxford.
- Rachel Freedman: admitted for AI PhD into Berkeley, Stanford, CMU and CS Master's at Cambridge; rejected from PhD at MIT. CV, Berkeley SOP.