A conversation with Professor Jared Kaplan, May 24, 2016

Participants

- Professor Jared Kaplan – Assistant Professor, Department of Physics and Astronomy, Johns Hopkins University
- Holden Karnofsky – Executive Director, Open Philanthropy Project

Note: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Professor Kaplan.

Summary

The Open Philanthropy Project spoke with Professor Kaplan of Johns Hopkins University about the possibility of encouraging people currently in the physics community to consider working on AI research, particularly with respect to reducing potential risks from advanced AI. Topics included potential benefits and downsides of outreach to early- vs. late-career physicists, and ideas for outreach, including holding a conference, offering research grants, and creating an undergraduate summer program.

Outreach to early- vs. late-career people in physics

Undergraduates and early-stage graduate students in physics are likely to be more open to considering career changes than more senior physicists. People earlier in their careers may also be more easily persuaded of the importance of AI research (including on reducing potential risks), since younger people tend to be more idealistic and open to new ideas. However, it is more difficult to identify talented people early in their careers. There are also incentives against making it known publicly that one is considering leaving physics, especially for people at a junior level.

Late-stage grad students or post-docs are less likely to be looking to change their career focus. However, there is likely to be more evidence as to their level of talent than for more junior people.

Potential motivations to change careers

People interested in switching careers away from academic physics are typically looking for one or more of the following:

- A better salary than is offered by a typical post-doc position. (Within AI research, some AI labs such as Google's would be competitive in this regard.)
- Fields with more practical applications (e.g. finance, computer science, engineering, applied science work within an industry).
- A role with a clearer or more direct positive impact on the world. Professor Kaplan expects it to be difficult to persuade physicists that AI research focused on potential risks (as opposed to, e.g., computer science or software engineering more generally) is a potentially highly impactful field.

Jobs for physicists in other fields

Physicists do not typically receive unsolicited job offers from other fields; those who leave physics tend to seek jobs proactively. However, there is no centralized, standard channel for searching for positions in other fields that are open to physicists. Professor Kaplan's impression is that physicists often find jobs in other fields through networking with their peers.

Corporations looking to hire physicists would likely hire grad students or recent post-docs and then use those hires' networks to find additional job candidates.

About ten years ago, finance firms were recruiting physicists and mathematicians. This may still be the case at a small number of firms.

Proposed AI safety conference

A two-day conference for physicists, where credible people in the field present on the importance of AI safety and raise the idea of AI safety research as a career path, could be beneficial.

Professor Kaplan thinks that grad students in particular would likely be willing to attend such a conference. He also thinks that more senior, established physicists could be persuaded to attend such a conference, though they would be unlikely to want to change career focus themselves. However, the conference could lead them to spread awareness of AI research as an option to more junior people, especially those who are looking to change careers.

Hosting a conference intended to recruit people from one field to another is, to Professor Kaplan's knowledge, a very uncommon practice and may be perceived as unusual.

Potential models for the conference

Two potential strategies for such a conference are:

1. Inviting a couple dozen junior physicists, along with five to ten more senior people in the field, and presenting the case for the importance of particular types of research to them all together.
2. Convening an initial conference for senior physicists aimed at persuading some of them of the importance of particular research, and then holding a second conference for junior people where those senior people help present on the topic and lend credibility to the issue.

Google "research days"

Google hosts "research days" to which it invites a few hundred scientists for presentations on a particular topic. If Google held a similar event that was either primarily or in part intended to promote AI research relevant to reducing potential risks, it might attract a large audience and be effective at recruitment.

Goals of the conference

Ideally, such a conference would create a general awareness that physicists who are interested in switching to machine learning (ML) research have the option of reaching out to the conference coordinators (e.g. the Open Philanthropy Project) for help making that transition. Types of assistance offered might include:

- Granting substantial fellowships for candidates who want to go into academic ML.
- Setting up interviews for jobs in ML.
- Connecting candidates with people working in the industry to learn more about the current state of AI research.
- Offering scholarships to support candidates while they study ML to get up-to-date on the field.

Risks

Ways in which this conference might fail to have an impact include:

1. The conference fails to attract a critical mass of attendees.
2. The conference does not cause attendees to take the research of interest seriously as an important field. (This risk might be reduced by having an initial conference for senior people and then having some of those people present at a subsequent larger conference for more junior people, as described above. Having representatives from well-known labs present could also lend the conference credibility.)

Potential obstacles to persuading physicists of the importance of AI research, particularly on reducing potential risks of advanced AI

People in physics largely do not consider this sort of work important by default. However, they typically do not have detailed enough knowledge of the field to engage with arguments for its importance at a technical level, and may not consider it a valuable use of their time to become more well-versed in the details of the field.

Scientists are accustomed to seeing particular subfields or research agendas advocated for as important on the basis of research that later turns out to be false or wrongly interpreted. Merely having a research agenda, a body of published papers, and the support of some credible figures are unlikely to be seen as compelling evidence of a field's importance.

Physicists might also be skeptical of an attempt to recruit them to work in a seemingly unrelated field, as it may raise the question of why recruitment is being done among physicists instead of people already working in computer science.

Another approach

Another potential route to recruiting researchers, including physicists, to AI safety research would be to offer substantial research grants and work to raise awareness that funding is available in this area.

Scientists may tend to apply for grants that would allow them to continue to work on essentially the same projects that they otherwise would have, and rarely change focus as the result of a grant. This dynamic makes it difficult to use grantmaking to influence what research gets done. However, Professor Kaplan thinks that a large enough amount of funding (i.e. tens or hundreds of millions of dollars) could have an influence and lead scientists to do research in the targeted area.

Jim Simons

Jim Simons of Renaissance Technologies donates roughly $10 million a year to theoretical physics research, which appears to have some impact on what research gets done (though, to Professor Kaplan's knowledge, Dr. Simons is not trying to influence the field in a particular direction).

Dr. Simons is currently offering grants of $2.5 million a year, over 4 to 7 years, to collaborations of professors who are aiming to create or reinvigorate specific subfields. These grants allow them to hire post-docs, organize summer schools, and generally attract people to their field by creating an impression that there are opportunities and funding available.

Dr. Simons also previously funded graduate fellowships but has stopped doing so.

Summer program

A summer program for rising junior and senior undergraduates studying computer science (or, potentially, math or physics), which offers a clear path to jobs in ML for interested students post-graduation, could be influential and relatively inexpensive. Ideally, the program would be prestigious enough to be valuable to students on their résumés.