# A conversation with Dr. Peter Eckersley and Dr. Jeremy Gillula, May 26, 2016

## Participants

- Dr. Peter Eckersley – Chief Computer Scientist, Electronic Frontier Foundation (EFF)
- Dr. Jeremy Gillula – Staff Technologist, EFF
- Holden Karnofsky – Executive Director, Open Philanthropy Project
- Helen Toner – Research Analyst, Open Philanthropy Project

**Note**: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Dr. Eckersley and Dr. Gillula.

## Summary

The Open Philanthropy Project spoke with Dr. Eckersley and Dr. Gillula of the Electronic Frontier Foundation (EFF) as part of an early investigation into how Open Philanthropy can support work on policy and strategy as they relate to long-term artificial intelligence (AI) outcomes. Conversation topics included AI computer security issues, international governance issues in AI, and the AI research field.

## Artificial intelligence (AI) computer security and operations security issues

The most important parameter affecting the security (specifically with respect to "snatching" a copy of it by an outside actor) of an AI program is the extent to which it communicates with the Internet. If it communicates freely and openly with the Internet, it is extremely difficult, and potentially impossible, to protect it from cyber theft. If communication with the Internet is constrained (for example, if only certain types of information can be communicated), the task is slightly less difficult. If the program is "air gapped" and does not communicate with the Internet, the task is easier still, but remains difficult against the most resourceful adversaries.

As cyber thefts aimed at major AI programs are likely to be committed by highly experienced teams, they may be very difficult to prevent.

### Preparing for AI security risks

*Timeline*

AI security risks can be difficult to prepare for, as it is uncertain when or how they will occur. Dr. Gillula believes that progress in the field will likely occur through a gradual, subtle process of continual software releases. As was the case during the Industrial Revolution, in the AI field, concepts and ideas are being spread and shared, and certain components are becoming interchangeable. While significant moments might be difficult to detect, Dr. Gillula believes that important technical developments will likely have a lead period of 1-2 years, and extreme scenarios should be considered and prepared for. It would likely take an AI firm approximately ten years to develop adequate security mechanisms to protect its programs from expert cyber thieves.

*Human resources*

Given the importance of the task, AI firms should hire top computer security experts to form a high-performing, in-house security team. It can be very challenging to hire high-quality computer security employees, and in larger organizations, ensure that all employees with access to critical systems are loyal. Firms like Google, Facebook, Microsoft, and possibly PayPal have been successful in the recruiting task, but many other corporations have not. Cultural fit is an important factor: for example, an AI firm's mission and workplace culture might be more attractive to top candidates than that of a typical banking institution. Institutions like the National Security Agency (NSA) are something of a middle case: they have strong recruiting pipelines, but may now be outcompeted by the likes of Google for many of the best hires.

Effective counter-espionage projects may require hiring former members of the intelligence community whose loyalty can be assured; this can be difficult for firms that are not part of or aligned with a nation state actor.

Google has hired numerous computer security experts who had previously worked for NSA. While Executive Order 12333 places some constraints on NSA's ability to spy on U.S. corporations, the resulting protection is relatively weak.

*Government computer security initiatives*

Dr. Eckersley believes that government computer security efforts are not fail-safe, citing a cyber theft incident in which the Chinese government stole U.S. weapons plans.

The field of AI security protection is likely to attract the interest of and potential assistance from multiple government agencies that might have differing objectives. It is important for external firms to collaborate with agencies that share their safety and security objectives, and ensure that clear jurisdictional boundaries are in place.

*Government intervention in AI progress*

In his role as Deputy U.S. Chief Technology Officer, Dr. Ed Felten has likely explored questions relating to the identification of potential AI-related national security risks and when and how government should intervene. He might be a valuable resource on these topics upon his departure from this position.

*Developing proprietary graphics processing unit (GPU) for deep learning neural networks*

One way to increase security in deep learning neural networks would be to develop proprietary graphics processing units (GPUs), as Google has done. Security properties would be built into a customized piece of hardware that would prevent bytes from being copied elsewhere. Parameters could be set to limit the types of data that can be processed, and alarms triggered when different patterns are detected. Algorithms could be designed to run much slower on non-proprietary hardware, but the proprietary hardware design could still be stolen and recreated.

While costly and potentially futile, the process of building proprietary hardware is likely a worthwhile learning endeavor. It has been done in the semi-conductor industry (by Intel and IBM), as well as the banking and film industries.

### Computer security concerns at the Cyber Grand Challenge

At the August 2016 DEF CON hacking convention in Las Vegas, the U.S.'s Defense Advanced Research Projects Agency (DARPA) will host the first round of the Cyber Grand Challenge (CGC), a hacking competition in which the contestants are computer programs. EFF has some concerns about the competition's security; for example, it is theoretically possible that successful entrant programs to such a contest could be repurposed for use in malware.

In 1988, a graduate student's project that aimed to gauge the size of the Internet (the "Morris worm") ended up crashing the Internet for several days. Nowadays, a piece of malware would need to attack hundreds of vulnerabilities in different systems to effect a crash of that magnitude, and at least conceptually, a successful CGC contestant program could find such vulnerabilities for itself.

## International governance issues in AI

There are several academic centers that might be well-placed to explore international governance issues in AI; examples include the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley and the Center for International Security and Cooperation (CISAC) at Stanford University, with which Dr. Eckersley has a technical affiliation. CISAC has expressed interest in exploring computer security issues.

A "constitution"-like agreement could help control the power of AI owners and manage scenarios in which multiple actors seek to use and/or adapt newly emerging AIs. It would set out the rights and responsibilities of relevant parties, including the AIs themselves, and would require significant creativity building on legal and political theory expertise to develop. Several scholars working at the intersection of law and AI who presented at a University of Washington workshop entitled "Artificial Intelligence: Law and Policy" might be interested in participating in the development of such an agreement. EFF also has the relevant expertise for this type of project.

### Artificial intelligence arms race

It would likely be very difficult to prevent an arms race to improve AI capabilities, primarily because the nature of such a race, and the types of AI that might be involved, are unclear in advance. Preparing for scenarios that do not end up occurring might be an inefficient use of resources.

## The AI research field

### Pace of AI research

An argument in favor of increasing the pace of AI research is that the hardware available to an AI system will increase over time; therefore, if artificial general intelligence (AGI) is developed sooner, it will have access to less "extra" hardware than if it is developed later. The amount of hardware beyond what an AGI is built on that it could potentially exploit, if developed, is referred to as "hardware overhang".

There are a number of other arguments to suggest that either slower or faster development would be more beneficial, so the answer is not clear.

### Open sharing of AI research

Dr. Eckersley believes that AI programs should only be shared openly if there is assurance that they will not be subject to strong evolutionary dynamics. Otherwise, there is a risk that an aggressive AI will become dominant and rapidly self-reproduce.

Dr. Gillula believes that, for this reason, AI programs that run only on proprietary hardware may pose less risk if shared openly than AI programs that run on commodity hardware.

OpenAI, an AI research organization, is interested in this question of encouraging open vs. closed AI work, and has asked for Dr. Eckersley's assistance in exploring it.

### Implications of different AI capabilities

Dr. Anders Sandberg and his colleagues at the Future of Humanity Institute (FHI) might be well placed to explore the capabilities and potential useful and/or important implications of different forms of AI.

### Developing the field of AI safety

The relative dearth of research activity in AI safety might be partly due to the absence of a robust research pipeline (for example, well-funded graduate or post-doctorate fellowships or professorships) in the field.

Academic and other research institutes that work in this area include:

- FHI at the University of Oxford
- Center for the Study of Existential Risk (CSER) at the University of Cambridge
- CLTC at the University of California, Berkeley
- Stanford Cyber Initiative and the Center for International Security and Cooperation (CISAC) at Stanford University
- Machine Intelligence Research Institute (MIRI)
- OpenAI (Dr. Eckersley has met with Ian Goodfellow of Open AI)
- Google DeepMind (Dr. Toby Ord has worked as a consultant for Google DeepMind)

Other pertinent academic domains might be international relations (for example, CISAC at Stanford University), law, and relevant interdisciplinary studies.

CLTC received significant funding for its cybersecurity work from the William and Flora Hewlett Foundation (Hewlett Foundation). As its efforts have helped build up the cybersecurity research field, the Hewlett Foundation might be able to provide guidance on doing this in the AI safety realm. Some CLTC researchers also received Future of Life Institute (FLI) grants. The FLI grant process has likely helped expand the pipeline of researchers working in this area.

If Dr. Eckersley were to publish a research paper in this area, he might do so under the purview of EFF or through another institution such as FHI. EFF currently considers AI safety to be an experimental area, but would be open to exploring certain issues on a contractual basis.

*All Open Philanthropy Project conversations are available at*
*http://www.openphilanthropy.org/research/conversations*