A conversation with Jacob Steinhardt, August 19, 2015

Participants

- Jacob Steinhardt – Graduate student in artificial intelligence, Stanford University and Science Advisor, Open Philanthropy Project
- Luke Muehlhauser – 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 Jacob Steinhardt.

Summary

The Open Philanthropy Project spoke with Mr. Steinhardt of Stanford University as part of its investigation into potential risks from advanced artificial intelligence (AI). The topic for this conversation was: "Is there computer security research which might not be important in the near term, and is thus not seeing much investment now, but which could be important when AI capabilities are substantially more advanced than they are today, and which could be productively studied today and might require decades of security research to address?"

Formal verification

Formal verification is the process of proving whether an algorithm satisfies particular requirements, or contracts.

The only contracts computer scientists are currently able to verify are specific, low-level requirements such as ensuring that an algorithm cannot access information defined as out of bounds or that a number stays in a specific range. Researchers lack the mathematical tools needed to specify complicated contracts with broad real-world applications.

It is already possible to "break" the abstractions used to specify these relatively straightforward contracts, via "side channel" attacks. As systems become more complicated and interact more with the real world, it becomes more difficult to specify what really matters in an abstraction, so these systems become vulnerable to an even wider range of attacks. This could pose security problems for complex near-future AIs.

According to Mr. Steinhardt, the question of how to explicitly and formally state complex real-world contracts is under-researched, but it is a topic researchers may be more interested in studying if they had better tools. If new approaches were developed, he believes researchers would likely conduct follow-up work.

Talent in the field

Another problem in Mr. Steinhardt's view is that, in part due to recent progress in machine learning and resulting excitement about the field, top talent tends to flow to machine learning more than it does to computer security. This is reflected both in the total number of researchers working in each field and in the career paths of the most talented researchers.

Defending against AI-guided cyber attacks

AIs may allow technically skilled bad actors to extend and magnify their influence. Two examples are given below.

Hacking more devices with greater control

AI systems may allow small groups of malicious hackers to extend their influence beyond what they are capable of individually. An intelligent software system capable of learning about a hacked system would allow hackers to control many more machines in a more flexible way than they are currently able to.

For example, it may be possible to use an AI to hack into a fleet of self-driving cars and instruct these cars to perform different, potentially coordinated, malicious acts, causing more damage than if a hacker were only able to issue a single command to all hacked vehicles.

Impersonating targets

If hackers are able to control more machines and use them to steal larger quantities of personal data, AIs may also help the hackers to better understand the data they acquire, enabling hackers to impersonate targets.

People to talk to

Relevant researchers and projects include:

- **Léon Bottou, Ph. D.** (Facebook AI Research)
- **Tom Mitchell, Ph. D.** (Chair, Machine Learning Department in the School of Computer Science, Carnegie Mellon University)
- **John Rushby, Ph.D.** (Principal Scientist, SRI International)
- **Future of Life Institute (FLI) grantees –** Some researchers funded by the recent FLI AI safety grants program, for example Alex Akin, may be working on or interested in working on the problem of formal verification for more complex systems.
- **Lean Theorem Prover** (Microsoft Research, Carnegie Mellon University)