A conversation with Peter Eckersley, May 27, 2015

Participants

- Dr. Peter Eckersley – Chief Computer Scientist, Electronic Frontier Foundation
- Dr. Nick Beckstead – Research Analyst, the 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.

Summary

The Open Philanthropy Project spoke with Dr. Eckersley of the Electronic Frontier Foundation (EFF) about artificial intelligence issues. Conversation topics spanned from present technology (how do machine learning and AI fit into the current picture of EFF's work?) to more speculative matters:

- EFF's work
- The current state of government surveillance
- Ways to mitigate the risks associated with foreseeable applications of AI, including surveillance and autonomous weapon systems
- Risks associated with brain emulations
- Ways to mitigate the risks associated with artificial general intelligence

Government surveillance

Machine learning techniques could make it easier for a surveillance agency to reconstruct the past history of an individual.

NSA surveillance

Currently, there is theoretically independent oversight of government surveillance programs in the U.S. However, this oversight is minimal and porous. National Security Agency (NSA) surveillance programs are overseen by a court established under the Foreign Intelligence Surveillance Act (FISA court). The FISA court is secret – prior to the Snowden leaks, the public could not read their judgments or review their decision-making. Historically, the FISA court has largely served as a rubber stamp, approving most NSA requests, and often having few details about the detailed nature of the programs it was approving.

XKeyscore

XKeyscore is a tool used by NSA analysts to search surveillance records. Its existence was revealed by Edward Snowden. From what is publicly known about XKeyscore, its operating procedure appears to be:

1. An analyst searches for key words.

2. XKeyscore returns a result, such as "170 email accounts match those key words; 56 of these belong to U.S. persons."
3. The analyst can immediately read emails from the non-U.S. accounts.
4. For the U.S. accounts there may well be a "Get a warrant" button available to analysts. To access U.S. email accounts, an analyst clicks this button and enters their reasons for accessing the account. (Dr. Eckersley emphasized that the details of this are speculative.)
5. There is probably an auditing procedure that ensures that these reasons are appropriate. However, intelligence analysts are probably able to come up with pretexts that would satisfy such an audit in a wide range of contexts.

Machine reading of email

The NSA / GCHQ's current policy appears to be to collect and store all email for at least some period of time. A warrant is required for human NSA analysts to read email that is clearly both sent and received by US persons. No warrant appears to have been required for the automated collection and storage of email, especially from foreign collection points (even though this includes a huge amount of US email). As far as NSA's internal legal opinions can be discerned from the outside, they may permit the use of machine learning techniques to identify which email accounts it should request a warrant for. EFF is pushing for a legal ruling to prevent machine learning being used this way.

Domestic, non-NSA surveillance

Courts oversee some types of non-NSA surveillance. For example, if a law enforcement agent wants to wiretap someone's phone, or read someone's email, a warrant is required. However, other types of surveillance can be conducted without a warrant.

The boundaries of the warrant requirement are the subject of ongoing dispute and revision by the courts. For instance, in the recent *Jones* case the Supreme Court established a warrant requirement for surveillance of location data. The Electronic Frontier Foundation (EFF) was involved in this litigation.

A lot of this fight centers around the erosion of fourth amendment protections, due to the wars on drug and terror. EFF uses impact litigation to defend and reinstate these protections where possible.

Third-party doctrine

The third-party doctrine holds that when personal documents are handed over to a third party (e.g. a bank), police can search them without a warrant. There is a question about whether this applies to electronic communication like email, or data held in the cloud. In a 2012 Supreme Court case (*United States v. Jones*), Justice Sonia Sotomayor suggested that perhaps the third-party doctrine was not appropriate for

the digital age and should be revisited. This issue is unresolved, and will probably be litigated in the future.

The law currently states emails are unable to be searched for the first 180 days, but can be searched after that (because they are "old records"). Dr. Eckersley believes that this portion of the law will at some point be overturned or removed by ECPA reform.

Encryption

Most of Dr. Eckersley's work is on encryption to help protect people from surveillance.

It's technically feasible to build computer architectures that are encrypted all the way down. In practice, no one builds systems like this, because they are less convenient to work with. While data is now often encrypted while being transferred, but is unencrypted when it is processed by cloud servers. It is more convenient to have cloud data be unencrypted, because the designs of server-side processing systems can be altered without having to reengineer the encryption.

Dr. Eckersley believes that strong encryption will begin to exist in a few specific areas. Text messaging data might become strongly encrypted. Messaging applications are relatively simple and easy to understand, making strong encryption easier to implement. Apple's iMessage application currently has strong security features that would make it very inconvenient for Apple to read a user's messages.

It is more difficult to apply strong encryption to more sophisticated applications like email or maps. Email needs to be searchable and filterable in various ways, which makes encryption more difficult.

Whistleblowers

The U.S. does not have a good system for protecting whistleblowers, especially whistleblowers with government security clearances. For example, the U.S. government aggressively pursues journalists and government employees that call attention to classified government information that arguably involves wrongdoing. The level of protection may vary by country – Iceland appears to have good whistleblower protections.

Work of the Electronic Frontier Foundation

EFF work divides into three main categories: political campaigning, impact litigation, and coding.

Activism

For example, encouraging people to call their member of congress and ask them to let the Patriot Act expire, or lobbying companies to change their practices on certain issues, or simply explaining complex technical and legal issues to a broad audience

in a comprehensible manner. The legislative component of this type of work is limited by EFF's 501(c)(3) status.

Impact litigation

EFF's traditional theory of change is based in impact litigation. EFF is "the ACLU of the Internet." This type of work comprises the largest part of EFF's budget.

EFF is engaged in a court fight over national security letters. National security letters are used by the Federal Bureau of Investigation to compel recipients to turn over information. National security letters are delivered with a gag order, which makes it a felony to tell anyone about receiving the letter. Recently, the Court of the Northern District of California ruled that national security letters are unconstitutional in the first round of litigation.

Code

Dr. Eckersley leads EFF's coding initiative. This initiative is relatively new – when Dr. Eckersley joined there were two people on the team. There are now approximately 10 people on this team (including contractors and part-time staffers). EFF has about 70 staffers in total.

The coding team produces software like *HTTPS Everywhere*, *Let's Encrypt*, and *Privacy Badger*. EFF software can be installed on web browsers and servers to make encryption and other privacy services very easy. Millions of people use this software.

Encryption campaign

In addition to long term progress on Web encryption, Dr. Eckersley believes EFF moved the needle on mail server to mail server encryption in the year after the Snowden leaks. About one third of mail providers previously encrypted their mail; this shifted to two thirds of mail providers encrypting their mail. To influence this, EFF published scorecards that showed which providers were encrypting appropriately and which were not. This encouraged engineers at those companies to implement encryption. Engineers at three or four of the top 20 email providers thanked EFF for its campaign, saying that the campaign enabled them to implement encryption at their companies.

(moving to somewhat more speculative topics around AI)

Mitigating the risks of autonomous weapon systems

Risks associated with autonomous weapons systems (e.g. military drones) could be mitigated by a number of strategies.

Shutdown codes

Shutdown codes (i.e. a mechanism that shuts the system down when passcode is entered) could increase the safety of autonomous weapons systems.

Shutdown codes are feasible if the drone is connected to a network. A network-connected drone can be programmed to shut down if a particular code is received over the network. The drone's operator could probably disable this function, unless the function was baked into the low-level hardware. Even then, the function could still be disabled by shorting a fuse, or wrapping part of the system in aluminum foil (to disable the antennae).

It's feasible for the shutdown code activation to require a key, or even a subset *N* of *M* keys. However, the keys could be stolen. Technological advances could make keys more secure (e.g. by putting the key on a special, secure USB stick).

Deciding who controls these devices is a difficult question. Managing the human decision-making around these devices is difficult.

A possible model for drone use policy is a landmine treaty. The Ottaway Treaty on landmines does not ban landmines categorically – it requires that landmines have remote controls (the US has not joined the treaty, but has taken the position that all of its mines should have end-of-life regulators). Landmines are a type of robotic weapon. Land mine policies are therefore a relevant model to study because they indicate what regulations are feasible (both nominally and in practice) for limiting the use of specific types of militarily tempting weapons. Landmine treaties have had some effect – probably about as effective as a good treaty could be -- though a significant number of countries are not parties.

Programming drones to be incapable of illegal activity

Some advocates of robotic and drone weapons have argued that they will be more law-abiding and ethical combatants than human soldiers.

It would be challenging to design drones that would be incapable of breaking laws because it can be hard to tell if the drone were breaking laws or not. Even humans can have significant difficulty distinguishing between legitimate and illegitimate targets in war. For example, in the *Collateral Murder* video (video of a July 12, 2007 U.S. airstrike which killed two Reuters journalists), the helicopter gunners were looking at blurry, long-distance video. The video showed pixelated people carrying camera equipment that looked somewhat like weapons.

In such a situation, it might be difficult for a drone directed by machine learning techniques to have significantly better discrimination than a human.

Human-in-the-loop requirement

People are more comfortable with drones when there is a "human in the loop", i.e. when a human does some part of the drone's decision-making.

Current U.S. military drones have a human in the loop – a person at a desk somewhere flies the drone and looks at incoming footage. The drone's software may propose targets, but the human must give approval before the drone attacks those targets.

Having a strong norm that military drones must have a human in the loop might limit the negative impact of future military drones. However, such a norm probably would not help if a government that did not respect the rule of law, such as in a coup. The human-in-the-loop norm would not be very useful if human operators did not cooperate with it (e.g. a drone pilot could accept all of the drone's attack proposals, thus not limiting the drone at all).

Unaccountable violence

Independent (non-state) actors could use autonomous weapons for attacks. The cost of carrying out a shooting is lowered when the shooter does not have to be present to pull the trigger. It is may be harder to identify the perpetrator of an attack when autonomous weapons are used.

Drone data transparency

Encouraging the publication of reports about current drone use would be beneficial. For example, the U.S. could release a report detailing:

- How many drones it has
- What it uses them for
- How many of its drones are armed, and how many rounds they carry
- What the armed drones attack

EFF has been pushing for drone transparency with respect to local governments and police departments within the U.S. It uses Freedom of Information Act requests to encourage the U.S. government to reveal information about its drones.

Machine learning techniques currently used by police

Some U.S. police departments currently use machine learning techniques to coordinate police officer deployment. These departments maintain databases with surveillance information, and use a machine learning program to make decisions about deploying police officers based off this information.

(Moving into more speculative topics, decades into the future)

Risks from brain emulation

Brain emulation is presently an extremely speculative technology with uncertain horizons for feasibility, but may be worth considering due to the potentially high stakes. There are several risks associated with brain emulations (computer models of human brains). These include:

- Humans with dangerous personality traits (including sociopath or other pathologies) who obtain power via brain emulation
- Situations in which some brain emulations have reason to fear other brain emulations (e.g. contests between groups of brain emulations). In these situations, emulations may have incentives to attack their competitors first. These attacks may escalate in unanticipated ways.
- Evolutionary selection of characteristics that do not align with current human values.
    - Emulations could be evolutionarily selected for efficiency, resulting in a world of "all work and no play."
    - Emulations could be evolutionarily selected for aggressiveness, which might have adverse outcomes.
    - In scenarios that initially have a variety of emulation types, emulations that try to reproduce as much as possible might become dominant due to evolutionary effects. In these scenarios, there may not be violence, but the selected type of emulation might come to own everything.

Inequality in societies with brain emulations

In contemporary Western society, power differences depend substantially on management and leadership ability in addition to wealth. A billionaire could decide to establish an organization, but if the billionaire is a poor leader, their organization will likely be unsuccessful. Conversely, if a middle-class person with good management skills decides to start a company or lead an organization, they can sometimes be very successful and thereby gain significant power.

In a society with brain emulations, this dynamic could significantly change. An emulation could create as many copies of itself as its resources permit. An emulation and its copies would all be aligned towards a common goal, so the burden of management is decreased. An emulation could immediately convert additional resources into additional capacity by making more copies of itself – a 1,000x increase in resources could translate into a 1,000x increase in capacity.

Artificial General Intelligence (AGI) interventions

Promoting emulation-oriented neuroscience research

Promoting emulation-oriented neuroscience research could reduce the risks of AGI. Moving the date of feasible brain emulation forward would reduce the potential CPU speed of the brain emulation when it occurred (i.e. if brain emulation occurred at a later date, the emulation would be able to run at a faster speed).

Neuroscience research contributes to understanding of brain data. Scanning technology research contributes to obtaining brain data. Both types of research will

likely be required to produce brain emulations. Differential development between the two could lead to differing scenarios (e.g. compare a scenario in which neuroscience research understands the brain well, but high-resolution brain scans don't exist, to a scenario in which high-resolution brain scans exist but the neuroscience is unable to explain the data).

There is a possibility that Moore's law ends before there is enough neuroscience understanding to run brain emulations.

Nature demonstrates that a functional human brain is possible within the volume of a human head. The biological brain is not a perfect analogy to silicon – clock speed is much slower, power requirements are much lower, and connectivity is much higher. There has not been an output of Moore's law that resembles the human brain. This is some evidence that even if Moore's law breaks down, there will probably be other directions to develop in.

Efforts to make AI development more open

Currently, most AI development occurs privately, at large companies. There is a risk that this type of development leads to the consolidation of power by a small group of people. Opening up AI development might mitigate this risk, though that would have to be weighed against any possible downsides of more open AI development.

All Open Philanthropy Project conversations are available at
http://www.givewell.org/conversations