A conversation with Dr. Peter Eckersley, August 4, 2015

Participants

- Peter Eckersley, Ph.D. – Chief Computer Scientist, Electronic Frontier Foundation
- Luke Muehlhauser – Research Analyst, Open Philanthropy Project

Summary

The Open Philanthropy Project spoke with Dr. Eckersley of the Electronic Frontier Foundation as part of its investigation into potential risks from advanced artificial intelligence (AI). The topic question 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?"

Under-investigated computer security problems

The development and widespread use of more advanced AI systems will likely introduce new computer security issues and exacerbate some existing problems. Given the pace of progress on AI capabilities commonly (but not universally) expected by AI experts, many of these problems are likely to be under-researched by computer security professionals relative to their likely importance and tractability.

Relying on AIs for security may introduce new vulnerabilities

As AI capabilities advance, parts of some systems' security may be implemented by complex AI methods that are based less on strict rules and careful reasoning (such as those of current encryption protocols) and more on the heuristics learned by the AI system. These AIs may be vulnerable to an AI-targeted equivalent of "social engineering," a type of attack on the human element of a secure system in which the attacker gains access to the system by exploiting human patterns of decision making and reasoning.

Analogous weaknesses of AI components in a security system may be difficult to identify in advance of an attack. It is difficult to ensure secure behavior even for systems in which all of the pieces are well understood, but unexpected behavior is especially likely for advanced AI systems, the behavior of which may be difficult to predict, especially in contexts that fail to resemble the AI system's training data.

Although there are some incentives to begin research related to this issue, Dr. Eckersley is uncertain to what extent relevant research is already in progress, but suspects that some of these avenues could benefit from additional resources. Possible areas of relevant research include:

- **Adversarial machine learning** – Incorporating adversarial training into the machine learning process could help mitigate the risk of AI-targeted attacks analogous to social engineering attacks on humans. This research would require identifying pathological cases that may not otherwise be part of the training population and using those cases in addition to the cases expected as part of the normal operation of the system.
- **Recognizing anomalies** – Another area of study that may be important is teaching AI systems to recognize not just the class of an input (e.g. a human face) but also when an input in a particular class is somewhat anomalous (e.g. a human face under duress).
- **Deploying AI security systems** – It will also be important to acquire sufficient experience deploying AI-using security systems to protect high value targets. Only when researchers have seen how these systems perform in real-world scenarios will they be able to understand whether AI defenses are robust or whether these systems are susceptible to frequent attacks.
- **The construction of "fuzzers"** – Fuzzers (tools that generate synthetic inputs to test the security properties of a system) are highly useful tools for testing traditional software systems, and there will likely be analogous tools developed to test AI systems, both of the "black box" and "white box" varieties. Black box fuzzers generate training inputs for a neural network while the network is monitored to ensure it does not return wildly incorrect answers. Using white box fuzzers, researchers are able to observe the system's changing internal parameter values and use those to identify the most pathological cases. Fuzzers have been deployed very successfully by corporations such as Google and Microsoft for improving the security of the systems they build, but the tools available to software developers at most institutions lag considerably behind the state of the art.

Some speculations about monitoring drones

Drones are sometimes nominated as a technology where radical AI policy concerns are imminent, though it is unclear how profoundly that is true. Their use by police forces extends existing concerns about surveillance cameras; their use in war extends existing concerns about aerial bombardment; their potential use by terrorists extends existing concerns about terrorism. While their increasing availability could potentially lower the costs of terrorism and other extreme antisocial behaviors, the extent to which these behaviors will increase is uncertain, and may be zero. Such acts will also most likely occur where they tend to occur

today (in areas with significant political upheaval), and many countries will likely deal with these acts in the same manner they currently do.

The increased use of drones could create a political imperative for new secure tracking systems. Mounting a gun on a drone is already relatively easy and is likely to become easier in the near future as drone range and controller mechanisms improve. If such a drone is used for an attack, the social response may be a backlash against the use of drones. But, if they are also widely used for commercial purposes and important sectors of the economy are dependent on drones, the response might instead be a call for an identification system that links drones back to human beings, similar to the license plate system used for vehicles. If there is pressure to develop such a system, substantial novel technical work would likely be required, as there are few examples of this type of distributed control architecture performing well in the face of adversity.

AIs that control critical infrastructure are a source of risk

A successful attack on a system that controls critical infrastructure can cause extensive damage very quickly. There are examples of this sort of attack on both the stock market and power grid (it took 27 days to repair the damage from a 2013 sniper attack on a Californian electrical substation), and the people who work in these fields are likely already engaged in explicit risk modeling.

However, these risks may increase as AIs become more sophisticated and their use becomes more widespread. If these systems rely on AI subsystems for dynamic stability management, then the AI itself becomes a significant source of risk.

In Dr. Eckersley's estimation, power grids and similar forms of complex civil infrastructure are a larger concern than the stock market, where there is a long history of using crude AIs in automated trading and adversarial behavior is the norm.

Obtaining confidence in the behavior of complex AIs will be difficult

Extraordinary engineering efforts are required to obtain sufficiently high confidence in the relatively simple software used today for autopilot on passenger jets and some passenger trains. Some experts think it may not be possible to obtain a high level of confidence in the more complex, adaptive and potentially self-modifying AI systems of the future.

This could have implications for how quickly complex AI systems can be deployed. According to Dr. Eckersley, it is possible to obtain some degree of confidence in the behavior of very complicated machine-learned systems, but it remains an open question whether this will need to be done on a case-by-case basis for individual systems or whether there is more general work that can be done in advance. If the former, it may be necessary to study an AI for years before its deployment to be

sufficiently confident in its behavior under adversity. Alternatively, if this work is generalizable, AIs could be deployed on a much shorter timeline (months rather than years).

One early step toward answering this question could be for computer security researchers to build systems by working with representative problem subspaces, then try to evaluate whether the security systems that work for these problem subspaces work for most cases or whether solutions will need to be invented problem by problem.

Monocultures are vulnerable to attack

Another open question in computer security research is whether a computer security monoculture is more vulnerable to attack than a diverse population of security approaches. A monoculture that has received strong security investment can be very difficult to break into, but if there is a break-in, all systems are compromised. For this reason, it is widely claimed that diversity may lead to increased security. Knowing whether this is actually the case may be valuable for making policy choices about secure AI deployment.

Hardware back-ups for critical systems are essential

Although the widespread use of more advanced AIs may exacerbate the need for hardware backups for critical systems, having redundant air-gapped versions of critical sub-systems is good design practice for secure systems regardless of the presence of AI systems.

Rebuilding hardware security structures

There is debate in the research community over whether it is possible to construct truly secure systems using currently available hardware. Some researchers (including Dr. Eckersley) believe that it will be necessary to re-build these structures from the ground up. There are multiple reasons for this:

- **Vulnerable side channels** – Cryptography depends on the absence of "side channels," but most devices have subtle vulnerabilities in their low-level hardware design, the result of a tradeoffs between security and efficiency. Modern computers are optimized to perform certain processes very efficiently, but by monitoring the outputs of those processes (including a device's time to execute code, power consumption, and radiofrequency emissions), hackers can gather information that can help them gain access to the system.
- **Outdated components** – Computers contain many smaller sub-computers that have historical designs dating from the 1980s and 1990s. The firmware

on these, and the components themselves, have never been rigorously re-engineered to withstand modern security threats.

Many critical systems rely on these insufficiently secure components. One example of a particularly vulnerable piece of hardware is the baseband chip that runs low-level control programs in mobile phones. These chips have well-documented bugs that can be remotely exploited to seize control of a device. Though some companies that produce these chips have increased their security expenditures, Dr. Eckersley is unaware of credible plans to engineer baseband chips that are secure, let alone secure in a way that users can verify.

Unless hardware structures are rebuilt from the ground up, AIs will operate on insecure hardware systems. AIs are likely already being developed behind closed doors as part of corporate research programs, but these researchers are typically incentivized to focus on being the first in their field to solve an AI problem, not on security engineering.

Dr. Eckersley believes that researchers should be working to anticipate the types of hardware that the most important AI systems will run on so that they can build the necessary hardware and operating systems in ways that allow them to be made highly secure.

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