CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Testimony before the Senate Armed Services Subcommittee on Cyber
Artificial Intelligence Applications to Operations in Cyberspace
Dr. Andrew Lohn
May 3, 2022

Chairman Manchin, Ranking Member Rounds and members of the Subcommittee, thank you for the opportunity to testify before you today. I am Andrew Lohn, Senior Fellow in the CyberAI Project of the Center for Security and Emerging Technology at Georgetown University. It is an honor to be here with Dr. Horvitz and Dr. Moore.

At the CyberAI project, we try to anticipate the impact of artificial intelligence and cybersecurity coming together. In these opening remarks I'd like to touch very briefly on three areas of that intersection: 1) How AI promises to improve cyber defenses, 2) How AI may improve offensive cyber operations, and 3) How AI itself is vulnerable.

Before I begin I would like to make clear that everything I am saying comes from an external vantage point. At CSET, we do not use classified sources and I do not have access to any private corporate data. Much of the cybersecurity world exists behind those closed doors, so there are surely capabilities or incidents that I am not aware of. However, much of it plays out in public, so we can try to extrapolate the future from the past.

Cyber Defense

AI for cyber defense is not a new concept. Spam and anti-phishing filters have been protecting users for many years. And AI has long been touted as a tool for companies that either hunt for malicious software or search for irregular behaviors that could indicate the presence of an intruder. Some of these techniques have become the foundations of modern cybersecurity while others are marketing hype. Sometimes it is difficult to tell the difference. In general, there is a back and forth where once an AI learns attacker tactics, those attackers adapt to evade that AI.

Cyber Offense

To date, those attacker tactics have not relied much on artificial intelligence. That is likely because so much has already been automated that humans only need to manage the attack. A human can select a computer script that scans the victim network and reveals possible targets. The human can then run another script that tries to exploit the vulnerabilities found by the first one. Then another script can enumerate the files and folders to encrypt or extract. The human only has to manage the system while computers already do most of the work.

That said, there are a few reasons to want the attack code to be able to make those decisions by itself. For example, the number of victims may be too large for humans to manage, or the targets may be difficult to communicate with over the internet. In 2015, when Russia first cut power to Ukraine, the hackers took over the mouse and had to manually select components of the grid to shut down. By the next year, they had developed new malware that was programmed with the ability to make some of those decisions without direct human involvement. The second version of that malware that was discovered last month is still being evaluated but appears to follow suit.

In addition to being able to operate where command and control might be difficult, an attacker may simply want to make decisions at machine speeds. In 2016, the year of the second power grid attack on Ukraine, DARPA hosted the Cyber Grand Challenge where fully automated systems competed to secure themselves while breaking into each other. These systems relied more on hardcoded rules than the advanced techniques we think of as AI today, but they showed some signs of promise. The winning automated system competed against some of the world’s top human teams the following day. Though it ultimately finished last, there were periods where it outscored some of the human teams, an impressive result in only its first year.

This was the first and last such challenge in the United States, but China was struck by the potential and has hosted at least seven of their own autonomous hacking challenges.[1] It is unclear how capable their systems are, but it is clear that both China and Russia are working to develop software that can discover vulnerabilities and, in some cases, is capable of running their cyber offensives more autonomously.

The threat extends beyond software that can autonomously find and exploit vulnerabilities. The human component is becoming more vulnerable. Humans are usually the weakest point in the security of a system, which is why 36 percent of intrusions involve phishing attacks.[2] Click rates have been falling for years but recent advances have made AI-generated text nearly as convincing as what humans can write. Combining that writing ability with the vast amounts of personal data on the internet provides a concerning potential for AI to make phishing campaigns even more effective than they already are.

Vulnerabilities of AI

[1] Dakota Cary, Robot Hacking Games, 2021.

[2] Verizon Data Breach Investigations Report, 2021.

Today's AI systems are technological marvels but they too are software complete with vulnerabilities of their own. They share some of the same vulnerabilities of more traditional software, but also introduce some new ones that can be very difficult to fix.

Most famously, it is easy for an attacker to change a few pixels in an image to make a detection system miss objects that it is looking for or to mistake objects in a scene for what the attacker wants them to see. Most strikingly, the attacker's manipulations can be so subtle that humans cannot tell the difference between the original and the doctored images.

It is easy to imagine these techniques being used to disguise parts of an invading force, or to direct autonomous search and destroy drones or coastal defense systems toward the wrong targets. It is even easier to envision digital decoys that overwhelm the system or its human operators. It is not clear yet how susceptible these systems are in the real world rather than just the laboratory setting, but we may find out soon, as many countries have become more keen to deploy autonomous military capabilities.

The United States is among those deploying autonomously capable systems, but our adversaries may not wait to subvert them. There are plenty of opportunities for interference throughout the design process. AI can be very expensive to train, so rather than starting from scratch, a system is often adapted from existing systems that may or may not be trustworthy. And the data used to train or adapt the systems may or may not be trustworthy too. It takes surprisingly few nefarious volunteers or low-paid online workers to corrupt a dataset in ways that give attackers a backdoor to control the model. Today most of these models and datasets are built and hosted by relatively trustworthy organizations such as those represented by Dr. Horvitz and Dr. Moore, but China in particular is making a push to provide more of these resources. If they succeed, then DoD would face an unwelcome decision between using the most capable systems or the most trustworthy ones.

Conclusion

I do not wish to overstate the impact of artificial intelligence on cyber security nor the severity of the vulnerabilities in AI. Cyber operations are still human-intensive both on offense and on defense. And there are few openly reported cases outside of a laboratory environment where AI algorithms were attacked directly. I only hope to alert you to the potential that is being developed. Our adversaries are highly capable and grow more emboldened every year. They have been developing increasingly autonomous attack software for years, and we should expect that those preparations will eventually come to fruition. Similarly, although we have seen only a few attacks directly on AI systems, the potential is no secret. Our adversaries are surely aware of the vulnerabilities and we should expect attacks as soon as AI systems prove their value on the battlefield.