Cyber Autonomy: Understanding and Mitigating the Risk of a Critical Operational Capability

Author: Dr. Misty Blowers, ICF

Abstract: The impact of a truly autonomous cyber capability could be profound. This paper intends to explore the benefits and drawbacks of autonomous cyber capabilities. It presents a method and emphasizes the importance of testing the robustness of the machine learning algorithms which form the basis for autonomous decision making. Robust assessment will help mitigate risks associated with nefarious actors compromising the learning systems. Considerations are also presented concerning the right level of human interaction and the importance of the human machine team. Finally, a brief discussion on how the latest uses of blockchain technologies could be used to improve the security and bound the behavior of cyber autonomous systems are discussed.

In July of 2017, the second-highest-ranking general in the U.S. military warned lawmakers in a US Senate Armed Services Committee Hearing against equipping the armed forces with autonomous weapons systems of which humans could lose control. Gen. Paul Selva advocated for keeping the "ethical rules of war" in place in discussing a directive that requires a human operator to be involved in the decision-making process when it comes to taking lives with autonomous weapons systems. (Brown, 2017) What lawmakers didn't discuss, however, is the role of autonomy in cyber operations. How much autonomy should we allow when considering incorporating intelligence into our cyber defensive and offensive strategies? Will we ever be able to respond to the speed of an adversary's cyber capabilities if we don't considered methods to combat autonomous cyber weapons? Most importantly, what will the world be like when autonomous defensive and offensive cyber capabilities start to fight each other? Are there scenarios where autonomous systems are a better option than using a human? How do you control the risk of autonomous cyber capabilities causing large-scale physical or economic damage that could trigger a kinetic attack response?

We have already seen what can happen when an autonomous cyber capability is unchecked. Consider how a cyber weapon like Stuxnet (intended for a single target system) found its way into industrial control systems throughout the world. First detected in July of 2010, it was estimated that the number of infected countries exceeded 115 only one month later. In August of 2010, Iran reported being the hardest hit, with the number of infections in the country at about 33,000. This is three times higher than the next most infected country, Indonesia, which reported nearly 10,000 compromised systems.

The reach of cyber autonomy is not even close to being fully understood. Autonomous agents are present in unmanned systems in space, air, land, and sea. They are increasingly becoming part of our command and control systems, logistics chains, and communications networks. There is a growing tendency for autonomy to be used for scenario planning and decision making, contingency management, fault detection and system health management. (Endsley, 2017) What do all of these applications have in common? The answer is in their interconnectivity to the world-wide-web which inherently makes them cyber capabilities. Consider the Dyn attack in October of 2016 that used specialized malware to target the systems of a major domain name service (DNS) provider and caused major internet websites to be unavailable for hours. This malware was specifically designed to exploit myriads of IoT-controlled devices (using their default passwords) and routed traffic from these captured devices

to the Dyn. The distributed and autonomous nature of this attack enabled the attackers to circumvent standard defense mechanisms. Some of these standard defense mechanisms include antiquated tools which are designed to do nothing more than detect a large volume of traffic from a single IP address. (Schneier, 2017) Even "air gapped" networks that are believed to be isolated and protected from the internet are vulnerable as the lines between electronic warfare and creative access strategies (like steganography, phishing schemes, and the use of social engineering) become blurred with the more traditional network access strategies. Unfortunately, the general public continues to remain blissfully confident in the systems in which they have come to rely for their security until hackers at security conferences, like DefCon, demonstrate just how many vulnerabilities are present in these systems.

Cyber autonomy offers tremendous potential for cyber network defense, especially when it is used to augment the role of the cyber operator. We have learned from both automation in manufacturing and on the battlefield that human augmentation can be designed to provide tremendous benefit and tends to be most beneficial when crafted to extend and complement human performance. Autonomy can extend human reach with enhanced perception, action, speed, persistence, size, and scale. It has the potential to expand the adaptive capacity of the warfighter and provide operations in contested environments. However, as speed of operations increases, the ability to maintain the human-machine synergy required for the success of the human-machine team will become more challenging and new information dissemination approaches will be required. There is a need for researchers to develop new ways to present the information from the autonomous agent to the human operator as the speed of the agent far out paces human information comprehension and decision-making. The humans must maintain the ability to oversee what they system is doing, intervene when needed, and maintain the ability to override the machine's actions when necessary. (Endsley, 2017) Policy makers on an international level need to take notice of the potential of these systems to have both remarkable benefits as well as devastating effects (both intended and unintended).

The US Department of Defense announced in 2016 that it planned to increase its investment into machine learning and autonomy. DepSecDef Bob Work noted that the US government should "Exploit all the advances in artificial intelligence and autonomy and insert them into DoD's battle networks to achieve a step increase in performance that the department believes will strengthen conventional deterrence." (Work, 2016) However, it seems like there is comparatively little research investigating the security risk associated with the machine learning mechanism associated with these autonomous systems. Not all machine learning methods are created equal, and it is critically important that decision-makers realize just how different approaches may respond to both intentional and unintentional corruption.

To illustrate this point further, consider the field of study called "adversarial learning". Adversarial learning is a robust assessment of machine learning-based approaches using the concept of adversarial drift: insert multiple points on decision boundary directly between what the model has characterized as normal operations and what the computer model has characterized as anomalous. In other words, teach a computer to learn with bad information, and it will learn to make bad decisions. A methodology to evaluate various machine learning algorithms was developed while examining and subsequently testing several Intrusion Detection and Intrusion Prevention Systems (IDS/IPS) with machine learning as the basis for their adaptation mechanism. (Blowers M. a.-1., 2014) (Nelson, 2014) The chart below illustrates the vulnerabilities of a few machine learning based approaches.

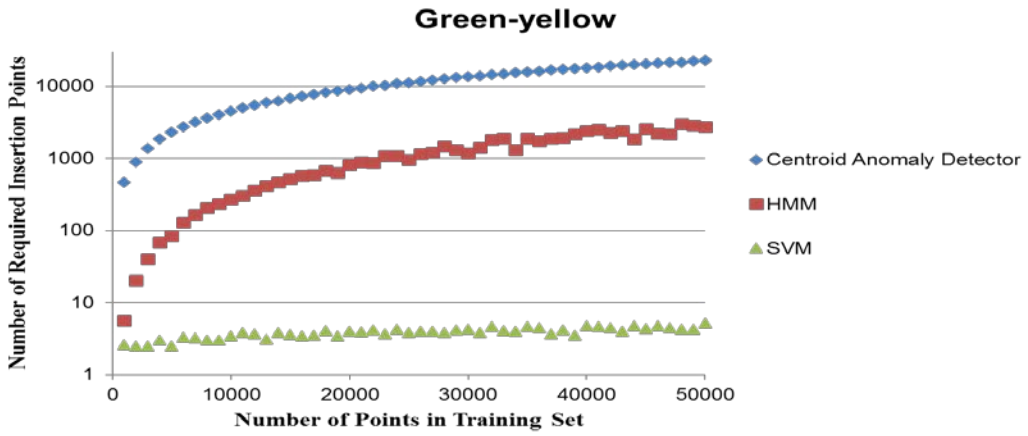

Figure 1: Evaluation of three machine learning methods susceptibility to a type of attack called "adversarial learning"

The study was targeting the machine learning failure points and decision boundaries where an adversary with ill intentions could influence accuracy of the machines representation of the environment in which it is operating. The chart shows the results of the assessment of intrusion detection systems equipped with 1) a Hidden Markov Model, 2) a Support Vector Machine, and 3) a Centroid Based Anomaly Detector. The support vector machine performed significantly worse than the other machine learning approaches, but this may not be the case for all implementations which leverage support vector machines. The point of this chart is to show how a machine learning assessment may performed, but it does not provide a full scope assessment of the hundreds of different machine learning approaches. What the chart does illustrates just how vulnerable a machine learning system can be to this type of corruption. It also illustrates the assertion that not all machine learning systems are created equal. Just as a human learning system can be corrupted, the learning mechanism of an artificial intelligence based system is just as vulnerable and some machine based learning system can be corrupted more easily than others. Just as a child can make bad decisions based upon bad behaviors it has learned from its parent, a computer can be fooled into making bad decisions that could have devastating effects and that effect the ability of the operator to maintain trust in the system.

Considering these factors, automation transparency is critical in establishing trust in autonomous cyber capabilities. The assumptions and goals of the autonomous system must be clearly understood. Testing of various machine learning systems must occur before they are deployed into a real world scenario. Autonomy which provides an intrinsic view into the decision boundaries and puts the sensitivity of the systems at the discretion of the user will allow the user to have oversight and prevent both intentional and unintentional compromise to the autonomous system.

As an additional layer of trust, distributed ledger technologies (blockchain) also may provide a means to secure cyber autonomy. Distributed ledger technologies offer a mechanism to ensure the information sent through a distributed network has not been compromised. If the blockchain infrastructure was used as a mechanism to bound the autonomous system and served to ensure that a set of rules of governance was followed regarding the autonomous systems evolution or adaptation to new environments, a trusted architecture would be realized. Features like time of creation, identity of the generating device, and authenticity of the underlying learning mechanism could be validated prior to introduction into the autonomous cyber agent's initial deployment, along with the bounding rules of adaptation.

Although there are many challenges presented in the use of cyber autonomy, we may be faced with the unpleasant reality in the near future that we MUST have such systems available at our disposal to combat an adversary that has developed an autonomous cyber capability that threatens our national security. The ideas presented here may help guide decision makers on a global scale when faced with such a dilemma. We need more tools to assess the robustness of the machine learning algorithms that guide the autonomous behavior and to protect the integrity of the data being fed to the systems. Creative solutions like the use of distributed ledger technologies and machine learning testing platforms discussed in this paper may help us maintain the control we need over an autonomous cyber capability so that we are acting responsibly. Finally, maintaining a human operator in the loop will continue to be necessary even when the cyber landscape created by the machines outpaces the speed of human response. There will be some level at which the human operator should and must maintain control to prevent an autonomous cyber action from causing wide-scale physical or economic damage. Innovative approaches will be required to maintain appropriate human oversight of a contested environment operating at machine-speed.

REFERENCES

Blowers, M. (2016). *THUTMOSE – INVESTIGATION OF MACHINE LEARNING-BASED.* Defense. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/1011870.pdf

Blowers, M. a.-1. (2014). Machine learning applied to cyber operations. In R. Pino, *Network Science and Cybersecurity.* (pp. 155-175). New York: Springer.

Brown, R. (2017, July 18). Retrieved from http://www.cnn.com/2017/07/18/politics/paul-selva-gary-peters-autonomous-weapons-killer-robots/

Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human factors, 59.1*, 5-27.

Nelson, K. G. (2014). Evaluating data distribution and drift vulnerabilities of machine learning algorithms in secure and adversarial environments. *SPIE Sensing Technology+ Applications. International Society for Optics and Photonics.*

Schneier, B. (2017, 4 12). *Lessons from the dyn ddos attack*. Retrieved 2017-4-12, from https://www.schneier.com/blog/

Work, B. (2016). Retrieved from http://www.defense.gov/News/Article/Article/991434/deputy-secretary-third-offset-strategy-bolsters-americas-military-deterrence