Policy Brief

Autonomous Cyber Defense

A Roadmap from Lab to Ops

Authors
Andrew Lohn
Anna Knack
Ant Burke
Krystal Jackson

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

CETaS Centre for Emerging Technology and Security

June 2023

Executive Summary

Given the immense economic and societal damage caused by cyberattacks and recent advances in artificial intelligence (AI), interest in the application of AI to enhance cyber defense has grown in recent years. Research is expanding on autonomous cyber defense that can not only detect threats but can engage in defense measures such as hardening or recovery. This report focuses on one promising approach to creating these autonomous cyber defense agents: reinforcement learning (RL).

There is no single agreed definition of autonomous cyber defense, but at its most basic level, these agents would complete some of the tasks of human cyber defenders by protecting networks and systems, detecting malicious activity and reacting to anomalous or malicious behavior, but at the speed of digital attacks.

This report presents a proposed definition for autonomous cyber defense, surveys the current state of autonomous cyber defense and the associated challenges that must be overcome for this technology to become a viable cybersecurity tool. There is no guarantee that autonomous cyber defense will succeed, but the technology is at a stage where policy support is needed to realize the potential benefits and help cyber defenders deal with the speed and uncertainty of modern cybersecurity operations.

RL is the leading AI approach to creating cyber defense agents, which are the core requirement of effective autonomous cyber defense. This technique increased in prominence in 2012 when RL agents first beat expert humans in simple Atari games. Building on that success, from 2015 and 2018, DeepMind built systems for far more challenging games, including Go and Chess, achieving unanticipated levels of success. Researchers flocked to RL, partly because of these successes, but also because of an open framework from OpenAI, which allowed creation of simple, simulated training environments or 'gyms.' The OpenAI gym format simplified research and development, and, in the last few years, cyber gyms have begun to appear that allow the training and creation of cyber defense agents. Even more recently, these gyms became part of an open cybersecurity competition titled Cyber Autonomy Gym for Experimentation (CAGE).

Our study is anchored on the potential for reinforcement learning (RL)-based AI agents to provide the autonomous capabilities required to fulfill some or all of the autonomous cyber defense concept. While the breadth of promising and relevant modeling approaches, techniques and technologies that relate to autonomous cyber defense is

large, our focus on RL is guided by the increased efforts in applied RL for cyber defense and the promising results RL has achieved in other problem domains.

While the technology central to autonomous cyber defense has advanced rapidly in the last decade, many challenges remain before systems can be deployed operationally. During the course of this research project, we interviewed government and non-government experts to identify the requirements for building and fielding trustworthy systems, which include:

- Adaptability - A potential autonomous cyber defense system will need to be future-proofed against changes in the cyber threat environment
- Auditability - Autonomous cyber defense systems must be able to generate logs and archive the agents' decisions and rationale in undertaking actions to enable review and audit, despite the operational tempo potentially exceeding human capacity. Audit logs can also be used to provide assurances that actions taken are lawful and proportionate, and adhere to agreed norms.
- Directability - Human operators will need to be able to redirect or terminate the system if needed.
- Observability - The system needs to provide human operators sufficient data capture and resolution to inform accurate, up to date situational awareness, and provide system performance metrics to support human oversight.
- Security - The autonomous cyber defense system and the agents within them all need to be secured against being leaked, stolen, or compromised.
- Transferability - Autonomous cyber defense agents will need to be deployable in real environments that do not exactly match the environment they were trained in.

To meet these requirements and continue progress, the fledgling field of autonomous cyber defense needs to be nurtured. RL has only recently started to take off for cybersecurity. Academic publications have surged in recent years and gyms for training cyber RL agents have begun to proliferate. However, capabilities remain rudimentary and incomplete compared to the more complex real-world network environments these agents will face. Sustained funding, coordinated effort to bolster simulation, emulation and evaluation tools, securing skilled personnel, and provisioning access to realistic data and infrastructure will help assure progress.

There is substantial potential for growth in autonomous cyber defense if technical challenges can be overcome. The existing agents and environments built for cyber defense currently consider fewer variables and possibilities than the more famous RL agents for playing board games like Go or video games like Atari or DOTA2. This means there is ample potential for increasingly intelligent agents; ones that can manage a larger number of possible defensive actions, and operate in more complex environments that require them to explore more situations. Our exploration of the technical challenges revealed that autonomous cyber defense is going to be a long-term ambition that can only be realized years into the future.

Recommendations

Despite significant progress in the autonomous cyber defense field, our study indicates that no autonomous cyber defense system has been deployed operationally. Given the present maturity of the current technology, we offer recommendations for developing these capabilities to mature the technology (See Chapter 5 for a full list of recommendations).

Invest in scaling up. The field can improve by making bigger and more realistic network simulations that incorporate more complex scenarios and attacker behaviors. Greater fidelity will lead to more capable cyber defense agents. In addition, releasing and maintaining tools such as gyms or trained agents can help attract academia or other researchers to do this work. Finally, sustained funding would also make it easier for researchers to align themselves to these projects.

Build and provide testing and training ranges. Larger and more complex agents will require more computationally intensive training and testing that could strain the resources of some researchers. Setting up and maintaining large computing systems is also a challenge, which requires talent that is hard to come by. Providing the requisite infrastructure, talent and funding resources – perhaps at a subsidized cost, could also help accelerate progress and provide continuity.

Coordinate data sharing. Policymakers across governments and industry have the power to release cyber data about networks that need to be defended and about threats that they are observing. These are all delicate issues that will require careful consideration, but to the extent that sharing data improves cybersecurity, all organizations stand to benefit.