CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Summary of Autonomous Cyber Defense - A Roadmap from Lab to Ops

The current AI-for-cybersecurity paradigm focuses on detection using automated tools, but has largely neglected holistic autonomous cyber defense (ACD) systems—ones that can act without human tasking. That is poised to change as tools are proliferating for training reinforcement learning (RL)-based AI agents to provide broader autonomous cybersecurity capabilities. The resulting agents are still rudimentary and publications are few, but the current barriers are surmountable and effective agents would be a substantial boon to society.

Specific visions of ACD vary among experts and in the literature.
Visions vary based on the range of tasks that agents should have available to them, the operational boundaries in which they operate, the role of humans in managing them, and the degree of autonomy they are granted. Current agents are limited in the complexity of the environment they can observe and the actions they can take, but there is substantial room for technological growth.

Recommendations

Given the promise and relative immaturity of the current technology, we offer recommendations for developing these capabilities:

- Nurture the field: Invest in scaling up the training gyms required to build and test cyber agents. Create specific data sharing programs for ACD. Host competitions to encourage innovation. Develop, attract, and retain AI and cybersecurity talent.
- Guide the field: Develop frameworks for understanding the risks and benefits of autonomous cyber defense. Determine thresholds for authorization of autonomous cyber defense agents. Establish and promote development priorities for autonomous cyber defense agents. Determine whether defensive agents can be developed without also developing offensive agents.

For more information:

- Download the report
- Contact Us: Drew Lohn, al1528@georgetown.edu