Machine Intelligence Cyberwar

Evaluating societal risks from military uses of Machine Intelligence Cyber Agents

Timothy Dubber and Prof. Seth Lazar

Australian National University

MINTLab Machine Intelligence and Normative Theory

Machine intelligence Cyber Agents (MICAs) are more feasible, and consequential, in the near term than other autonomous weapons.

Cyberspace is a more constrained domain than the physical world. Standardised protocols define its terrain, and sensory input is inherently machine-readable. Thanks to this constrained nature, automatic programs, like malware worms, can already navigate the live and changing cyberspace environment. Thus, developing autonomous agents in cyberspace is less technically burdensome.

Furthermore, machine intelligence is already being employed across the full spectrum of cyber-attack operations, from reconnaissance to command and control. The only major hurdle remains simulating cyber actors' "actions-on-the-objective"—turning access and opportunity into a broader operational plan.

Moreover, because the cyber domain impinges more directly on broader social, economic and information ecosystems, MICAs have the potential to impact civilian daily life outside of wartime.

Thus, MICAs are more likely to be fully realised before other domains of warfare. Their development will impact commercial, government, civil society, and critical infrastructure networks. Their potential and near-term feasibility will probably spark an arms race with state, commercial and cybercriminal actors seeking to build offensive and defensive MICAs.

The development of MICAs poses catastrophic risks to the public internet. If one escapes human control, it would be nearly impossible to eradicate.

Autonomous cyber agent models selected for resilience will likely look to self-replication and redundancy of model locations to ensure agent survival during conflict. If a model is self-transparent or has access to the code base for other autonomous cyber actors, it may attempt to 'seed' autonomous cyber actors outside their original home networks.

Ongoing efforts to shrink machine intelligence models increase the number of locations a model could deploy itself—from large data centres and high-performance computing clusters to, potentially, personal and office networks. This widens the scope of autonomous cyber actors' hiding places. Furthermore, digital trends of increasing internetworking, the proliferation of IoT and persistently low levels of cyber hygiene ensure such models will continue to have many places to hide.

With such preconditions, a MICA could go rogue, escaping into the public Internet, no longer responding to its sponsor's tasking, and selecting targets based on the model's internal logic. These models could dwell semi-dormant across global networks and telecommunications infrastructure, occasionally re-emerging to attack networks or infrastructure. Remediation of such an infestation would require an almost complete rebuild of the Internet.

We should mitigate risk through safeguards and limit MICA proliferation, but we also need to invest in defensive cyber agents.

A pure abolition approach to MICAs is unlikely to work. Previous cybersecurity agreements—like the 2015 US-China cybersecurity compact—have failed due to state intransigence and the inherent opacity and secrecy of cyber programs. And the presence of active spoilers, like cyber criminals, further undermines the feasibility of such initiatives.

However, limited agreement to restrict proliferation and avoid complete devolution of targeting critical infrastructure to MICAs could slow the emergence of a worst-case scenario. Furthermore, researchers' precautions, like restricting MICAs from their own code and building in fail-safes, reduce the risk of breakout scenarios.

But, considering the likelihood of an eventual breakout, efforts should also be made now to develop the defensive MICAs necessary to overcome and potential risk posed by offensive MICA models.

Select bibliography

Buchanan, Ben. (2020) *The Hacker and the State*, Cambridge, Massachusetts: Harvard University Press

Clarke, Richard A.; Robert K. Knake, (2019), *The Fifth Domain: Defending Our Country, Our Companies, and Ourselves in the Age of Cyber Threats*, Penguin Press

Kott, Alexander S.. (2018) *Intelligent Autonomous Agents are Key to Cyber Defense of the Future Army Networks*, US Army Research Laboratory

Lohn, Andrew J., et al., (2023) *Policy Paper: Autonomous Cyber Defense*, Centre for Emerging Technology and Security,

Margulies, Peter S. (2020) "Autonomous Weapons in the Cyber Domain: Balancing Proportionality and the Need for Speed" in *International Law Studies*, Vol. 96

Oesch, Sean. et al., (2024) "The Path to Autonomous Cyberdefense" in *IEEE Security & Privacy*, no. 01