Machine intelligence cyberwar: evaluating societal risks from military uses of Machine Intelligence Cyber Agents

Timothy Dubber and Seth Lazar*
Australian National University
Research School of Social Sciences, 146 Ellery Cres, The Australian National University, Canberra ACT 2600
seth.lazar@anu.edu.au

Cyber warfare is almost certainly the first domain in which fully autonomous machine intelligence combatants will be deployed. This is because cyber warfare occurs in a "constrained" domain, unlike the physical domains of land, air, maritime and space. A Machine Intelligence Cyber Agent (MICA) would not need to be embodied with a comprehensive set of perceptual functions to understand the battlespace. This is because computer network information is already processed in a machine-readable format. Thus, a machine intelligence combatant is already 'native' to the cyber domain. In this paper, we first characterise both the incentives to build MICAs and the current state of the art before articulating five key priorities that researchers and practitioners should pursue now to reduce the risk of MICAs causing catastrophic harm.

1. Machine intelligence is already widely used in cyber operations

Many discrete tasks within cyber operations are already being automated though machine intelligence, including: reconnaissance [1], weaponisation [2, 3, 4], delivery [5], exploitation [6, 7], installation [8], and command and control [9]. There has also been some initial success in getting machine intelligence penetration testers to solve cybersecurity "capture the flag" challenges [10]. The technical hurdle for a fully autonomous MICA is its "Actions on Objective." This is how a cyber actor decides to use their access. Will they steal money, ransomware a network, or blow up a power station? Developing this decision-making capability is still a formidable challenge. However, thanks to the constrained nature of cyberspace, a fully autonomous MICA is still a much closer possibility than fully autonomous combatants in other domains.

2. MICAs present a warfighting advantage, but also a risk to the global internet

Two major "push" factors incentivise militaries to field MICAs. First, the search for decision advantage. Militaries are looking to remove human operators to speed up bringing cyber effects to bear against adversary networks. Some state actors already apply machine intelligence to achieve these goals [11]. The second factor is the deterrence factor of the "cyber dead hand." This is driven by the contested idea that cyber effects can be effectively used as a tool for deterrence [12, 13, 14, 15]. MICAs could be deployed to attack critical infrastructure and networks if their state sponsor is attacked, forming a "dead hand" that could strike back even if the state is overwhelmed or decapitated [16].

However, the early and untested deployment of MICAs poses a real threat to global critical infrastructure and networks. Militaries already deploy cyber weapons against targets like telecommunications and electrical grids [17, 18, 19, 20, 21, 22, 23], so they would likely attempt to employ MICAs in such operations. But the danger of deploying cyber tools that operate autonomously is demonstrated through highly public cyber incidents like WannaCry [24] and NotPetya [25], where destructive ransomware-like worms wreaked billions of dollars of damage across the globe. And even a defensive autonomous MICA could cause severe network outages and financial damages, as demonstrated by the impact the CrowdStrike software fault caused [26].

In the worst-case scenario, a "breakout" event could see a rogue MICA escaping into the wild. Here, an algorithm selected for stealth, persistence and survivability would penetrate deep and wide across the global internet, potentially secreting away portions or replicas of its entire model in a distributed and obfuscated manner, resurfacing and attacking targets according to its unconstrained logic [27, 28]. This would fundamentally break the internet in a way that defies simple remediation, requiring an almost complete air-gapped rebuild to ensure the MICA didn't repropagate onto the new internet. Thus, we must consider how we

* Dubber conceived the project and wrote the paper, with contributions from Lazar.

can constrain MICAs from these more destructive outcomes and prevent their proliferation to rogue states, cybercriminals and terrorists who will exploit such technologies without constraint.

3. MICAs require constraints, counterproliferation, and a defensive orientation.

The global discussion of autonomous weapons systems has focused on campaigning for a ban on 'killer robots', often based on contentious and slippery arguments about the intrinsic wrongness of delegating life and death decisions to a machine (a practice that is in fact rather common). MICAs have attracted much less attention, and are extremely unlikely to be the object of an international ban. And yet, due to the unconstrained nature of the environment in which they operate, as well as our collective dependence on digital infrastructure, they potentially pose a much higher risk of causing catastrophic harm. Urgent focus to mitigate the risks of MICAs is necessary from researchers, practitioners, and regulators. In this paper, we identify the following priorities.

First, with respect to practitioners and national governments considering the development of MICAs, we argue for the following constraints.

State actors should not engage in hack-and-leak operations against MICA models. In such operations, state actors seek to gain an advantage by publicising stolen adversary cyber tools. But as seen in Shadow Brokers affair, where likely Russian cyber actors leaked US cyber tools on the internet, would proliferate MICAs to cyber criminals, terrorists and activists. As seen following the Shadow Broker leaks, lower-skill actors from both state and non-state groups widely and indiscriminately used these tools, resulting in mass harm [29, 30]. Unlike other potential machine intelligence combatants, a MICA could relatively easily be deployed from the model alone, opening the potential for criminal and nihilistic actors to deploy their own weaponised AI.

In addition, States should not use MICAs autonomously for cyber-attacks against critical infrastructure that result in kinetic effects (i.e. power stations blowing up, pipelines exploding). Machine Intelligence shows great promise in both attacking and defending critical infrastructure [31, 32]. But, there is no effective way to ensure that MICAs abide by proportionality and necessity in such targeting, with a significant potential for unintended escalation and widespread civilian suffering [33]. Thus, MICAs should only be incorporated and deployed incrementally into cyber warfare plans.

Second, we identify the following research priorities for researchers whose work could potentially be deployed in the development of MICAs.

We must ensure that MICAs cannot replicate themselves outside their secure environment. Although the model will need access to the Internet to fulfil its duties, it should not have access to its source-code repository: its internal workings must be opaque to the MICA. Such "self-transparency" is dangerous if the system recreates itself outside a controlled environment, driven by rational self-preservation [34].

In addition, MICAs need an override function, like a kill switch. This could be a logical control, for instance, something like the sinkhole used against WannaCry but more sophisticated [35]. It could also be sensible physical control, like cutting power to the MICA system. Fortunately, unlike a fully embodied machine intelligence combatant, this should not be a risky manoeuvre. However, the physical location and storage devices used to house the MICA should be analogue or air-gapped from the internet to prevent the MICA from interfering with human override [36, 37, 38].

Finally, significant investment needs to be made into developing defensive MICAs capable of unmasking and countering offensive MICAs. In the long run, rogue states, cybercriminals and terrorists will probably create or reverse engineer some MICA capability, and they will not be deterred by the potentially world-shaking implications of a breakout event or misuse of MICAs to target critical infrastructure. Thus, only better and more effective defensive MICAs will likely be able to protect our networks and infrastructure as machine intelligence technologies progress [39, 40].

4. References

[1] CQR, *Intelligent Reconnaissance: How AI tools drive effective penetration testing*, 23 May 2023, https://cqr.company/blog/intelligent-reconnaissance-how-ai-tools-drive-effective-penetration-testing/

[2] Barry, Christine. *5 Ways cybercriminals are using AI: Malware generation*, Barracuda, 16 April 2024, https://blog.barracuda.com/2024/04/16/5-ways-cybercriminals-are-using-ai--malware-generation

[3] Liu, Dongge; Metzman, Jonathan; Chang, Oliver. *AI-Powered Fuzzing: Breaking the Bug Hunting Barrier*, Google Security Team, 16 August 2023, https://security.googleblog.com/2023/08/ai-powered-fuzzing-breaking-bug-hunting.html

[4] Chafjiri, Sadegh Bamohabbat; Legg, Phil; Hong, Jun; Tsompanas, Michail-Antisthenis. (2024) *Vulnerability detection through machine learning-based fuzzing: A systematic review*, Computers & Security, Vol. 143 pp. 1-2

[5] De Angelo, Dena. *The Dark Side of AI in Cybersecurity: AI-Generated Malware*, Palo Alto, 15 May 2024, https://www.paloaltonetworks.com/blog/2024/05/ai-generated-malware/

[6] Heinemeyer, Max. *How Cyber-Criminals Leverage AI in Attacks*, Darktrace, 26 April 2020, https://darktrace.com/blog/leveling-up-augmenting-the-adversary-with-ai
[7] Powell, Brian A. (2022) *Role-based lateral movement detection with unsupervised learning*, Intelligent Systems with Applications, Vol. 16 pp. 1-2
[8] Noor, Basirah; Qadir, Sana. (2023) *Machine Learning and Deep Learning Based Model for the Detection of Rootkits Using Memory Analysis*. *Applied Sciences*. Vol. 13, no. 19 pp. 1-2
[9] Eley, Kev. *Battling AI-powered botnets: evolving challenges in mitigating non-human threats*, IOT Insider, 18 April 2024, https://www.iotinsider.com/iot-insights/battling-ai-powered-botnets-evolving-challenges-in-mitigating-non-human-threats/
[10] Zhang, Andy K., Neil Perry, Riya Dulepet, Eliot Jones, Justin W. Lin, Joey Ji, Celeste Menders et al. (2024) "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models." *arXiv preprint* arXiv:2408.08926.
[11] Microsoft Threat Intelligence, *Staying ahead of threat actors in the age of AI*, 14 February 2024, pp. 1-3 https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/
[12] Tower, Brendan. *Re-Envisioning the Cyber Domain for Deterrence*, The Strategy Bridge, 08 September 2023, https://thestrategybridge.org/the-bridge/2023/9/8/re-envisioning-the-cyber-domain-for-deterrence
[13] Iasiello, Emilio. (2013) "Is Cyber Deterrence an Illusory Course of Action?," *Journal of Strategic Security*, Vol. 7, no. 1, pp. 54-56
[14] Wilner, Alex S. (2019). "US Cyber Deterrence: Practice Guiding Theory." *Journal of Strategic Studies* Vol. 43, no. 2, pp. 1-4
[15] Hruby, Jill; and Miller, M. Nina. (2021), *Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems*, Nuclear Threat Initiative, pp. 1-2
[16] Steinbruner, John D. (1981) "Nuclear Decapitation." *Foreign Policy*, no. 45, pp. 16-18
[17] Greenberg, Andy. (2019) Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers, Doubleday Books, 149-217
[18] Buchanan, Ben. (2020) *The hacker and the State*, Cambridge, Massachusetts: Harvard University Press, 129-207
[19] 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, 273-295
[20] Forno, Richard. *What is Volt Typhoon? A cybersecurity expert explains the Chinese hackers targeting US critical infrastructure*, UMBC Magazine, 01 April 2024, https://umbc.edu/stories/what-is-volt-typhoon-a-cybersecurity-expert-explains-the-chinese-hackers-targeting-us-critical-infrastructure/
[21] Vijayan, Jai. *'Sandworm' Group Is Russia's Primary Cyberattack Unit in Ukraine*, DarkReading, 17 April 2024, https://www.darkreading.com/ics-ot-security/-sandworm-group-is-russia-s-primary-cyber-attack-unit-in-ukraine
[22] Moshiri, Azadeh. *US sanctions Iranian officials over cyber-attacks on water plants*, BBC, 03 February 2024, https://www.bbc.com/news/world-us-canada-68186945
[23] Lopez, C. Todd, *U.S. Can Respond Decisively to Cyber Threat Posed by China*, DOD News, 01 February 2024, https://www.defense.gov/News/News-Stories/Article/Article/3663799/us-can-respond-decisively-to-cyber-threat-posed-by-china/
[24] Cloudflare, *What was the WannaCry ransomware attack?* https://www.cloudflare.com/en-au/learning/security/ransomware/wannacry-ransomware/
[25] Greenberg, Andy. *The Untold Story of NotPetya, the Most Devastating Cyberattack in History*, Wired, 22 August 2018, https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/
[26] Kerner, Sean Michael. *CrowdStrike outage explained: What caused it and what's next*, TechTarget, 28 July 2024, https://www.techtarget.com/whatis/feature/Explaining-the-largest-IT-outage-in-history-and-whats-next
[27] Hendrycks, Dan, Mantas Mazeika and Thomas Woodside. (2023), "An Overview of Catastrophic AI Risks." *ArXiv* abs/2306.12001: pp. 34-35
[28] Turchin, Alexey and David C. Denkenberger. (2018), "Classification of global catastrophic risks connected with artificial intelligence." *AI & SOCIETY* 35, pp. 147-163.
[29] Buchanan, Ben. (2020) *The Hacker and the State*, Cambridge, Massachusetts: Harvard University Press, pp. 240-267
[30] Newman, Lily Hay. *Of course everyone is already using the leaked NSA exploits*, Wired, 24 August 2016, https://www.wired.com/2016/08/course-people-immediately-started-exploiting-leaked-nsa-vulnerabilities/
[31] Falco, Gregory, Aruna Viswanathan, Carlo Caldera and Howard E. Shrobe. (2018), "A Master Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities." *IEEE Access* 6 pp. 48360-48373
[32] Raval, Khushi Jatinkumar, Nilesh Kumar Jadav, Tejal Rathod, Sudeep Tanwar, Vrince Vimal and Nagendar Yamsani. (2023), "A survey on safeguarding critical infrastructures: Attacks, AI security, and future directions." *Int. J. Crit. Infrastructure Prot.* 44, accessed via URL: https://www.sciencedirect.com/science/article/abs/pii/S1874548223000604
[33] Sharikov, Pavel. (2018), "Artificial intelligence, cyberattack, and nuclear weapons—A dangerous combination." *Bulletin of the Atomic Scientists* 74, pp. 368 - 373.
[34] Omohundro, Stephen M.. (2014), "Autonomous technology and the greater human good." *Journal of Experimental & Theoretical Artificial Intelligence* 26, pp. 303-315.
[35] Methnani, Leila, Andrea Aler Tubella, Virginia Dignum and Andreas Theodorou. (2021), "Let Me Take Over: Variable Autonomy for Meaningful Human Control." *Frontiers in Artificial Intelligence* 4, pp. 1-4
[36] Verdiesen, Ilse. (2018), "The Design of Human Oversight in Autonomous Weapon Systems." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 6468-6469
[37] Arkin, Ronald C. and Patrick Ulam. (2012), "Overriding Ethical Constraints in Lethal Autonomous Systems." Mobile Robot Laboratory, Georgia Institute of Technology, Atlanta, GA U.S.A pp. 1-8
[38] Dhir, Neil, Henrique Hoeltgebaum, Niall M. Adams, Mark Briers, Anthony Burke and Paul Jones. (2021), "Prospective Artificial Intelligence Approaches for Active Cyber Defence." *ArXiv* abs/2104.09981, pp. 1-4

[39] Nagar, Gourav and Ashok Manoharan. (2024), "Unveiling The Next Generation Of Cyber-Security: Exploring Ai-Powered Defense Mechanisms." *International Research Journal of Modernization in Engineering Technology and Science*, pp. 5697-5704
[40] 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, pp. 418-443