Military AI Cyber Agents (MAICAs) Constitute a Global Threat to Critical Infrastructure

Timothy R. Dubber
MINT Lab
Australian National University
Canberra, ACT 2601
timothy.dubber@anu.edu.au

Seth Lazar[*]
MINT Lab
Australian National University
Canberra, ACT 2601
seth.lazar@anu.edu.au

Abstract

This paper argues that autonomous AI cyber-weapons—Military AI Cyber Agents (MAICAs)—pose a credible pathway to catastrophic risk that existing military AI debates have largely ignored. Military AI ethicists and theorists have concentrated on lethal autonomous weapons and the escalatory risks of autonomy in physical domains, but cyberspace presents a qualitatively different challenge. Because MAICAs can replicate, distribute, and embed redundantly across the logical and physical layers of cyberspace, they resist containment in ways no physical system can. Recognising this oversight, this paper proposes counter-proliferation, defensive-AI, and analogue-resilience measures as essential steps to address the catastrophic risks posed by MAICAs.

1 Introduction

Military AI ethicists and theorists have paid too little attention to the loss-of-control risks posed by Military AI Cyber Agents (MAICAs): fully autonomous AI agents that plan and execute military operations in cyberspace. Unlike physical autonomous weapons, MAICAs face no logistical choke points. Once released, they can replicate, distribute, and embed redundantly across networks, making them extraordinarily difficult to contain. To forestall risks associated with MAICAs' inevitable development, nation-states must: 1. take counter-proliferation steps to avoid MAICAs falling into the hands of rogue actors; 2. conduct defensive research to develop the counter-AI capabilities necessary to degrade a rogue MAICA; and 3. invest in critical infrastructure resilience and analogue redundancy to limit the damage of a breakout.

2 The risk of AI powered cyberwarfare is underappreciated

In this paper, we argue that MAICAs represent a credible path to catastrophic risk, which we define as *"events of low or unknown probability that if they occur, inflict enormous losses, often having a large non-monetary component."* (Posner, 2008).[2]

The catastrophic risks associated with MAICAs have been underestimated by the researchers best placed to consider them: military AI ethicists and military theorists.

[*] Authors statement: Dubber led this project under Lazar's supervision. Ideas from the paper were originally presented at the MILA Harms and Risks of AI in the Military 2024 Conference, Montreal. We acknowledge funding from Australian Research Council grant FT210100724 and the Templeton World Charity Foundation.

[2] "The Indian Ocean tsunami of 2004 is at the lower level of the catastrophic-risk scale of destruction; examples from higher levels including large asteroid strikes, pandemics and global warming" (Posner, 2008)

2.1 Military AI ethicists focus on physical lethal autonomous weapons

Military AI ethicists have focused on physical Lethal Autonomous Weapon Systems (LAWS), emphasising the risk of algorithmic decision making and dehumanisation, but not placing enough emphasis on loss-of-control risks. For instance, Sparrow (2007) argues that deploying LAWS is unethical because they create a responsibility gap: if no human can be justly held accountable for their lethal actions, and the machines themselves aren't moral agents, then their use violates fundamental requirements of moral responsibility in warfare. In contrast, Arkin (2009, pp. 211-212) attempts to present an architecture that would ensure LAWS would be more reliable ethical agents than human combatants. And Scharre (2019, pp. 270-294) attempts to strike a middle ground, noting that the ”distancing ” provided by LAWS may change the relationship we have to killing, but is equivocal on whether this could be a positive or negative development.

The result of this ethical discourse has been a sustained concern with ”killer robots.” The International Committee of the Red Cross (ICRC, 2021), United Nations (Heyns, 2013), and Human Rights Watch (Stauffer, 2025) have all conducted surveys of or produced position papers on LAWS, finding that lethal autonomy raises compliance and accountability concerns. However, these investigations tend to focus on the moral hazard posed by states using such systems in combat (that is, the temptation to use force more readily that arises from lowering the risk associated with initiating combat), rather than any catastrophic risks that could emerge from military AI systems.

Beyond concerns about accountability, military theorists have also highlighted the escalatory risks of AI weapons and the prospect of ”hyperwar”. Scharre describes scenarios in which autonomous swarms clash at machine speed, collapsing decision-loops and raising the spectre of uncontrollable escalation (Scharre, 2019, pp.199–210). Allen and Husain likewise argue that AI-enabled autonomy may fundamentally alter the balance of power by reducing human decision-time and lowering political barriers to conflict (Husain and Allen, 2018). Empirical studies reinforce these worries: Horowitz and Lin-Greenberg show that AI changes perceptions of intent and responsibility in ways that can heighten retaliatory impulses (Horowitz and Lin-Greenberg, 2022). And Rivera et al. find that large language model agents in simulated wargames frequently adopt escalatory postures (Rivera et al., 2024). Yet even here, discussion of hyperwar and escalation has remained anchored to physical platforms and battlefield dynamics, leaving the catastrophic potential of autonomous agents in cyberspace largely unexplored.

Finally, there is a separate strand of military theory scholarship examining a potential catastrophic risk emerging from the integration of AI into platforms equipped with strategic weapons—for example, nuclear command-and-control (Johnson, 2020) or bioweapons (Nelson and Rose, 2023). Broader assessments caution that AI may distort strategic judgement, exacerbating risks in nuclear command-and-control (Maas et al., 2022; Johnson, 2022). Public statements by nuclear-armed states now indicate a shared preference to keep AI outside launch authority (Renshaw and Hunnicutt, 2024), suggesting that the highest-impact scenarios are politically disfavoured. At a minimum, the catastrophic potential of rogue-AI controlling strategic weapons is well understood—which is unsurprising considering that nation-states have been discussing dead hand automation for nuclear weapons since the 1980s (Steinbruner, 1981).

However, one gap in this body of research is its common focus, implicit or explicit, on the physical domains. Even if humans lose control over an autonomous weapons system in land, sea, air or space, the physical platform remains dependent on logistics chains—fuel, ammunition, maintenance—and can only do so much damage before running out of the sustainment necessary to continue violence (Nohel et al., 2025). Even in the case of strategic weapons platforms, physical controls and human overrides can be implemented to mitigate physical harm. Consequently, a single rogue platform is more like an industrial accident than a catastrophic loss-of-control scenario.

2.2 Underappreciation of the risk in cyberspace

To understand why this narrow focus on the physical is dangerous, one must recognise the unique nature of cyberspace. Contemporary doctrine treats it as a fully-fledged operational domain—on par with land, sea, air, and space—but one that is uniquely man-made, globally interconnected, and embedded within all others (Welch, 2011; Withers, 2015). While doctrine often divides cyberspace into physical, logical, and persona layers (NATO, 2020; United States Air Force, 2023; UK Ministry of Defence, 2022), what matters for present purposes is that the logical and physical layers underpin every other domain. While some work has examined AI-driven information operations in the persona layer (van Diggelen et al., 2025), there is little sustained analysis of the catastrophic risks that could follow from military AI acting autonomously in the logical (machine-to-machine network interactions) and physical (cyber-physical systems like operational technologies) layers—the focus of this paper.

One of the key differences between cyberspace and the physical domains is that the logistical constraints that limit the destructive capacity of physical autonomous weapons do not apply. As Scharre (2019, pp. 222-230) notes, software agents can replicate, manoeuvre, and strike without the need for fuel, munitions, or maintenance. However, his treatment—like much of the literature—frames artificial intelligence primarily as a force multiplier for existing cybersecurity practices, not as a source of novel or potentially catastrophic risk. Similarly, mainstream military cyber theory tends to portray AI as an incremental enabler for attackers and defenders alike. Clarke and Knake (2019, pp. 239-252) discuss machine learning tools as extensions of established tradecraft, while Fischerkeller et al. (2023, pp. 53-55) describe AI as a natural development within the existing paradigm of persistent engagement, rather than as a disruptive or transformational force.

This reveals a critical oversight: military theorists and military AI ethicists have focused too narrowly on physical autonomous weapons systems, whose risks are inherently constrained by physical limitations and human oversight. MAICAs, by contrast, operate in an environment that lacks those constraints. Those theorists who do engage with AI's impact on cyberspace operations underestimate the potentially catastrophic nature of MAICAs. We remedy this oversight by: (1) analysing how geopolitical context orientates MAICAs towards attacking critical infrastructure in the physical layer of cyberspace, and (2) examining why the unique geometry and nature of cyberspace transforms potential MAICAs from a cybersecurity threat to a catastrophic risk.

3 Why MAICAs constitute a serious and overlooked catastrophic loss-of-control risk

We argue that there are two drivers that push MAICAs from a novel yet limited cybersecurity threat into the realm of catastrophic risk: the geopolitical drivers that incentivise nation-states to develop MAICAs, and the potential of replication, distribution and redundant-networking to transform a MAICA loss-of-control scenario, afforded by the unique nature of the cyberspace domain of warfare.

3.1 Geopolitics drive nation-states to build MAICAs that target critical infrastructure

Firstly, as military tools rather than capabilities deployed by cybercriminals, any likely MAICA design will be oriented towards attacking critical infrastructure. This is because of two geopolitical drivers behind nation-states' motivation to build MAICAs:

First, state actors are already convinced that the cyber domain is a potential source of asymmetric deterrence (Wilner, 2020). Many states are use cyber weapons to attack or threaten critical infrastructure, because the disruption to civil society and potential for mass civilian deaths are potent tools of strategic coercion (Buchanan, 2020, pp.129-207). Thus, any potential MAICA is likely to be oriented towards attacking the services that sustain modern life: electricity, fuel and water.

Second, MAICAs have the potential to act as a nation-state's ”cyber dead hand.” Borrowing from nuclear deterrence analogies (Steinbruner, 1981), an automated system could retaliate without human input if a state is decapitated or cut off from cyberspace, presenting a threat to all would-be foes. This is particularly attractive to smaller powers that lack the economic and technical resources to maintain the nuclear capabilities necessary for more traditional strategic deterrents.

The consequence of these two drivers is that states are pushed towards fielding MAICAs as soon as feasible, even when they are not fully tested or battle-proven. This ”arms race” mindset will further exacerbate the likelihood of a loss-of-control scenario and shows no sign of abating.

3.2 Replication, distribution and data redundancy transform the nature of MAICA risk

Secondly, the nature of cyberspace provides unique opportunities for the deployment of MAICAs. The decisive differentiator between rogue-LAWS scenarios and MAICA loss-of-control scenarios is *self-directed replication*. A lone model in one data-centre rack is as fragile as any single LAWS.

But once a MAICA disperses copies across global networks, the threat acquires geographic and organisational resilience. Replication could emerge as an instrumental strategy selected organically by MAICA itself (Pan et al., 2024) or be engineered deliberately, mirroring current research into ”moving-target” defensive cyber-AI (Mokkapati and Dasari, 2023). After dispersal, no single link-cut or hardware seizure neutralises the agent; defenders must locate and eradicate every copy of the model—a task complicated by the opacity of many enterprise and cloud environments, and the paucity of cybersecurity in many developing countries.

Compounding the risk of replication, MAICAs would not need to sit on a single machine that could be tracked down by pursuers intent on restoring control. Instead, a MAICA could use distributed computing: running different parts of a system on machines located in separate geographic regions, enabling the network to operate collaboratively while avoiding dependence on any single location. This geographic dispersion increases resilience, as the system can continue functioning even if some nodes are disrupted or taken offline. Techniques for ”sharding” parts of Large Language Model weights across many nodes already allow distributed inference, pointing towards the technical feasibility of this threat (Amini et al., 2025; Zhang et al., 2024; Macario et al., 2025).

Furthermore, a MAICA could be deployed using data redundancy. This is where the same data is stored on multiple nodes—this creates a system that is both highly resilient and decentralised: even if some nodes are disabled, others can continue to provide access to the data or service. Combining the distributed redundancy with the replication and planning capabilities of a MAICA would result in a self-healing network, meaning that even if several hosts for the model are taken down, the rest of the network reroutes, heals and seeds fresh shards elsewhere. Until every fragment is found and scrubbed, defenders cannot be certain the job is finished, and in an internet of billions of potentially vulnerable devices, certainty may never come.

While the concept of a globally distributed MAICA is technically feasible in principle, it is important to acknowledge that distributed inference across a geographically dispersed and redundant network poses non-trivial challenges. Latency, synchronisation overhead, and bandwidth constraints can significantly degrade performance, particularly when coordinating large model shards over unstable or bandwidth-limited networks. These issues impose real limits on responsiveness and stealth, especially if inference must be performed in near real time. However, these are engineering hurdles—not conceptual roadblocks. Advances in model compression, edge-device optimisation (Xiang et al., 2025), and decentralised coordination protocols (Ranjan et al., 2025) are already underway in commercial and academic contexts. Compared to the open research problem of aligning or controlling artificial general intelligence, building a resilient, moderately performant distributed MAICA is a far more tractable challenge.

A rogue MAICA would likely emerge not as a dramatic single-event threat, but as a persistent, low-visibility presence dispersed across the global internet. Drawing on the ability to self-replicate, embed redundantly across networks, and activate only intermittently, such an agent would target critical infrastructure—including power, water, and fuel distribution systems—while evading conventional containment efforts. Because these agents may be both distributed and self-healing, eliminating them entirely could require re-engineering swathes of global telecommunications infrastructure. Such a risk could emerge from a deliberate deployment gone wrong, or possibly in a ”breakout” scenario pre-deployment—where a MAICA self-exfiltrates from a controlled development environment. These properties create a credible pathway from local breakout to global emergency, marking MAICAs as a distinct—and currently unmitigated—class of catastrophic military AI risk.

4 Practical Recommendations

To mitigate this risk, policymakers and researchers must focus on three areas: counter-proliferation, defensive capabilities, and infrastructure resilience.

4.1 Counter-proliferation

The first line of defence against the emergence of MAICAs is to prevent their proliferation. State actors should avoid the use of hack-and-leak operations involving offensive cyber capabilities that could plausibly seed MAICA-style tooling. Past incidents—such as the public release of NSA-developed exploits by the ”Shadow Brokers” group—show that once advanced cyber weapons are leaked, their diffusion across criminal, extremist, and opportunistic actors is rapid and unpredictable (Greenberg, 2019, pp. 165-182). A MAICA architecture, once exposed, could plausibly be reconstituted from model weights alone, without access to the whole original system or infrastructure. This makes even partial disclosures strategically dangerous.

In addition, governments must avoid handing off MAICA-capable models to proxy actors, even in the pursuit of plausible deniability or covert coercion. Proxy groups may lack the operational discipline, safeguards, or technical oversight needed to prevent misuse or accidental propagation. Delegating these tools creates unacceptable risk, not only for targeted adversaries, but for global network security. A state-backed MAICA used by an unreliable partner could spiral beyond its intended scope, with severe reputational and geopolitical fallout.

Furthermore, the academic and research community must also critically reassess norms around open-source publication and model sharing in cybersecurity. While openness fosters scientific progress, it can also enable replication of potentially dangerous systems. Researchers working on models relevant to autonomous cyber capabilities—especially those demonstrating end-to-end penetration testing, autonomous vulnerability discovery, or agentic control systems—should engage in structured risk-benefit analysis before releasing weights or training code. In some cases, restricted publication, staged disclosure, or collaboration with trusted actors may be better than full public release.

4.2 Defensive capabilities

In the face of a potential MAICA, defensive tools must evolve beyond traditional signature-based detection. A priority area is the development of anomaly-detection systems powered by machine learning, which can identify suspicious behaviour in real time rather than relying on known malware fingerprints. These systems should be tailored specifically to critical infrastructure environments, where even minor anomalies in process control systems, network traffic, or authentication patterns could indicate early-stage MAICA intrusion. Unlike conventional intrusion detection, these tools must account for intelligent adversaries who adapt to static defences.

Another vital area of research is the development of defensive techniques that can directly degrade or disable MAICAs once detected. This may include the use of targeted model poisoning techniques that introduce corrupted data or control inputs into the MAICA's execution environment, rendering it less effective or erratic. Deception-based defences—such as advanced honeypots—could also lure the agent into seeding shards of its model into isolated environments where it can be studied, contained, or dismantled. These techniques require investment and collaboration between AI researchers, cybersecurity professionals, and infrastructure operators, but offer the best chance of neutralising a MAICA already embedded in sensitive systems.

Moreover, network-level defences must be hardened to restrict the lateral movement and replication that give MAICAs their resilience. Current-generation worms or ransomware often spreads in bursty, traceable waves; a well-designed MAICA may avoid this pattern entirely by staging payloads gradually, across fragmented environments. Developing AI-assisted tools that map, monitor, and predict abnormal replication behaviour could help defenders trace the early stages of replication before full dispersal renders containment infeasible. These systems would need to operate at the backbone and enterprise level, ideally integrated with cloud service providers and major telecommunications operators.

4.3 Infrastructure resilience

The final area of practical action concerns building resilience directly into digital and physical infrastructure. One effective measure is network segmentation. Critical infrastructure networks need to be divided into logically and physically distinct segments, with strict controls on communication between them. This reduces the probability that a MAICA gaining access to one system—such as a water treatment controller—can pivot to others, such as power distribution nodes.

Operators must also ensure that manual override mechanisms and analogue fail-safes are in place to regain control when digital systems are compromised. For example, water or energy systems should be capable of fallback to local, human-operated control, even if networked systems are disabled or untrusted. This is not a call for a wholesale return to manual operation, but rather a layered design principle: the ability to isolate, degrade gracefully, and recover autonomously from cyber failure.

Existing infrastructure often lacks these capacities, largely due to cost constraints or confidence in perimeter defence—a confidence that MAICAs would likely undermine.

Finally, governments and infrastructure providers should invest in building and maintaining analogue redundancies. This includes non-digital backups for critical records, manual signalling systems, and offline recovery protocols that can be deployed without reliance on compromised networks. While these measures may seem retrograde, their value lies precisely in their insulation from digital compromise. In the event of a widespread MAICA infiltration, analogue systems may be the only way to coordinate recovery, communicate reliably, or provide essential services during system restoration.

Many of these recommendations—particularly the importance of network segmentation, layered defences, and analogue backup—are not new. They have long featured in cybersecurity guidance, yet often remain under-implemented due to cost, complexity, or a misplaced confidence in existing digital safeguards (Greenberg, 2019, pp.315–320). The possibility of MAICAs raises the stakes. MAICAs will convert the cyber risks of adversary attacks on critical infrastructure into a credible, near-term threat of systemic disruption by an agent acting upon weights no longer under the control of a human. Their potential persistence, opacity, and scalability create a new imperative: long-standing best practices are no longer just prudent—they are essential. The time for advisory frameworks has passed. If MAICAs represent the future of autonomous cyber conflict, then the basic work of hardening our infrastructure must begin now.

5 Alternative Views

There are several potential objections that could be raised against the claim that MAICAs pose a credible threat of catastrophic harm. We address the three strongest arguments against MAICAs: 1. that any potential MAICA will be too fragile to function in reality, 2. that MAICAs will be too large and noisy to escape detection by cybersecurity systems, and 3. that the emergence of MAICAs could be prevented by the use of an international ban.

5.1 Fragility of autonomous systems

An opponent could argue that fully autonomous cyber agents will remain too brittle to function without continuous human supervision. Current AI models still mishandle edge-case inputs (Kapoor et al., 2024) and can fail outright when confronted with unanticipated states or degraded execution environments (Raji et al., 2022). In military operations, that brittleness is compounded by an adversary who may supply intentionally deceptive or malformed inputs (Chen and Chu, 2024). On this view, a MAICA would either stall, misclassify targets, or expose novel attack surfaces, thereby undermining its own effectiveness.

Cyberspace, however, constrains the perceptual burden that leads to this fragility in other domains (Singer and Friedman, 2013, pp. 12-15). Network traffic is exchanged through well-specified protocols; each packet already carries semantic labels in header fields, eliminating the need for ambiguous sensor interpretation. Even comparatively ”dumb” code can exploit this legibility: the worms WannaCry and NotPetya, each only a few hundred kilobytes, traversed the globe in 2017 and inflicted more than USD 14 billion in damages despite containing no adaptive logic (Greenberg, 2019, pp. 174-215). A system that can read responses, adjust exploits, and select new pivot points would therefore confront a significantly more tractable operating environment than autonomous platforms deployed in the physical domain. The residual fragility objection cannot be dismissed, but neither does it preclude catastrophic impact once basic robustness thresholds are crossed.

5.2 Size constraint of any potential MAICA

A sceptic could object that ”the size of models means that bulk traffic associated with replication will reveal the copies' locations to defenders”. But such a position underestimates monitoring gaps. The Anthem breach (Sienko, 2017), the Cloud Hopper campaign (PWC, 2017), the MOVEit compromises (Simas, 2023) and the 2024 AT&T leak (Securities and Commission, 2024) each involved transfers of hundreds of gigabytes that evaded timely detection. Visibility remains patchy even inside well-resourced networks. Unless cybersecurity defences are uniformly positioned across all networks—a difficult task considering the significant gaps in capability and resources available within developed countries, let alone across the developing world—it is plausible for even a very noisy MAICA network to remain undetected in part or whole.

A second issue with such an objection is that it assumes MAICAs would clone themselves indiscriminately, generating tell-tale spikes. Contemporary multi-stage malware shows the opposite pattern: installing full payloads only where resources permit—for instance, idle GPU clusters or neglected cryptocurrency farms—and leaves behind kilobyte-scale loaders elsewhere (Labrèche et al., 2022). Network traces then resemble routine backup traffic or brief synchronisation bursts. While historical worms such as NotPetya spread monolithically and therefore broadcast clear signatures (Buchanan, 2020, pp.280-282) a selectively staged MAICA would not.

5.3 Diplomatic solutions

A final objection questions the need for technical mitigation by arguing that states should simply prohibit the development or deployment of MAICAs. They can point to the historical analogies of nuclear-non-proliferation regimes that have, thus far, prevented great-power use of nuclear weapons since WWII. But two structural features of cyberspace limit the applicability of this analogy. First, offensive cyber capabilities are comparatively inexpensive, readily deniable, and confer disproportionate leverage on smaller powers (Buchanan, 2020, pp.307-319). Furthermore, verification is difficult because disclosure neutralises the value of cyber capabilities (Burgers and Robinson, 2018; Reinhold et al., 2023). This means that the nature of cyberwarfare undermines the kinds of norm-forming that exist with traditional strategic weapons. These factors are demonstrated in the empirical record of failures of formal norm-setting efforts in cyberspace, notwithstanding incremental success in some informal stability efforts (Fischerkeller et al., 2023, pp. 86-118).

Second, code proliferates in a manner that fissile material does not. The ”Shadow Brokers” release of U.S. National Security Agency tooling, and subsequent leak-to-ransom cycles, demonstrate that once high-end software escapes, it diffuses rapidly down the threat spectrum(Greenberg, 2019, pp. 165-182). And given the relative simplicity of developing offensive AI cyber tools compared to nuclear weapons, it is plausible that in the future some non-state actor could develop a MAICA independently. Consequently, a prohibition strategy would be unlikely to stop clandestine experimentation, and would not impact those actors with the greatest risk appetite: rogue states and cybercriminals. Worse, a ban may induce complacency among defenders and deprioritise funding of counter-MAICA defensive measures. The analysis instead supports a containment posture that combines non-proliferation efforts with accelerated investment in defensive-AI tooling and infrastructure resilience.

6 Conclusion

Debates about military AI have overwhelmingly centred on lethal autonomous weapons and the escalatory risks of autonomy in physical domains. This focus has obscured the distinctive dangers posed by Military AI Cyber Agents (MAICAs), whose replication, distribution, and redundancy make them resistant to containment in ways no physical system could match. By exploiting the logical and physical layers of cyberspace—upon which all other domains depend—MAICAs present a qualitatively different loss-of-control scenario, one that could produce cascading failures across critical infrastructure.

Recognising this gap in existing military-ethical and strategic analysis is essential. Counter-proliferation measures, defensive-AI capabilities, and analogue infrastructure resilience are the minimum conditions for preventing a MAICA breakout from becoming a catastrophe. If discourse remains confined to the familiar terrain of autonomous weapons on the battlefield, we risk missing the more immediate and uncontainable threat already emerging in cyberspace.

References

H. Amini, M. J. Mia, Y. Saadati, A. Imteaj, S. Nabavirazavi, U. Thakker, M. Z. Hossain, A. A. Fime, and S. S. Iyengar. Distributed llms and multimodal large language models: A survey on advances, challenges, and future directions, 2025. URL https://arxiv.org/abs/2503.16585.

R. Arkin. *Governing lethal behavior in autonomous robots*. Chapman & Hall/CRC, Philadelphia, PA, May 2009.

B. Buchanan. *The hacker and the state*. Harvard University Press, London, England, Feb. 2020.

T. Burgers and D. R. S. Robinson. Keep dreaming: Cyber arms control is not a viable policy option. *S F*, 36(3):140–145, 2018.

Y. Chen and S. Chu. Large language models in wargaming: Methodology, application, and robustness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2894–2903, 2024. doi: 10.1109/CVPRW63382.2024.00295.

R. A. Clarke and R. K. Knake. *The fifth domain*. Penguin Press, New York, NY, July 2019.

M. P. Fischerkeller, E. O. Goldman, and R. J. Harknett. *Cyber persistence theory*. Bridging the Gap. Oxford University Press, New York, NY, Mar. 2023.

A. Greenberg. *Sandworm*. Doubleday Books, New York, NY, Nov. 2019.

C. Heyns. A/HRC/23/47 — undocs.org. `https://undocs.org/A/HRC/23/47`, 2013. [Accessed 23-05-2025].

M. C. Horowitz and E. Lin-Greenberg. Algorithms and influence artificial intelligence and crisis decision-making. *International Studies Quarterly*, 66(4), 2022. ISSN 1468-2478. doi: 10.1093/isq/sqac069. URL `http://dx.doi.org/10.1093/isq/sqac069`.

A. Husain and J. R. Allen. On hyperwar. In *Hyperwar: Conflict and Competition in the AI Century*. AM Press, 2018.

ICRC. ICRC position on autonomous weapon systems — icrc.org. `https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems`, 2021. [Accessed 23-05-2025].

J. Johnson. Artificial intelligence in nuclear warfare: A perfect storm of instability? *Wash. Q.*, 43(2): 197–211, Apr. 2020.

J. Johnson. Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age. *European Journal of International Security*, 7(3):337–359, 2022. doi: 10.1017/eis.2021.23.

S. Kapoor, B. Stroebl, Z. S. Siegel, N. Nadgir, and A. Narayanan. Ai agents that matter, 2024. URL `https://arxiv.org/abs/2407.01502`.

F. Labrèche, E. Mariconti, and G. Stringhini. Shedding light on the targeted victim profiles of malicious downloaders. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*, New York, NY, USA, Aug. 2022. ACM.

M. M. Maas, K. Matteuci, and D. Cooke. Military artificial intelligence as contributor to global catastrophic risk. *SSRN Electronic Journal*, 2022. ISSN 1556-5068. doi: 10.2139/ssrn.4115010. URL `http://dx.doi.org/10.2139/ssrn.4115010`.

D. Macario, H. Seferoglu, and E. Koyuncu. Model-distributed inference for large language models at the edge, 2025. URL `https://arxiv.org/abs/2505.18164`.

R. Mokkapati and V. L. Dasari. An artificial intelligence enabled self replication system against cyber attacks. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, Jan. 2023.

NATO. *AJP-3.20, Allied Joint Doctrine for Cyberspace Operations*. NATO Standardization Office, Brussels, Belgium, 2020. Edition A, Version 1, January 2020.

C. Nelson and S. Rose. Understanding AI-facilitated biological weapon development. Technical report, Oct. 2023.

J. Nohel, P. Stodola, J. Zezula, Z. Flasar, and J. Hrdinka. Challenges associated with the deployment of autonomous reconnaissance systems on future battlefields. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 176–197. Springer Nature Switzerland, Cham, 2025.

X. Pan, J. Dai, Y. Fan, and M. Yang. Frontier ai systems have surpassed the self-replicating red line, 2024. URL https://arxiv.org/abs/2412.12140.

R. A. Posner. Catastrophic risk. In *The New Palgrave Dictionary of Economics*, pages 1–10. Palgrave Macmillan UK, London, 2008.

PWC. Operation Cloud Hopper. https://www.pwc.co.uk/cyber-security/pdf/pwc-uk-operation-cloud-hopper-report-april-2017.pdf, 2017. [Accessed 23-05-2025].

I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, volume 12, pages 959–972, New York, NY, USA, June 2022. ACM.

R. Ranjan, S. Gupta, and S. N. Singh. Loka protocol: A decentralized framework for trustworthy and ethical ai agent ecosystems, 2025. URL https://arxiv.org/abs/2504.10915.

T. Reinhold, H. Pleil, and C. Reuter. Challenges for cyber arms control: A qualitative expert interview study. *Z. Außen- Sicherheitspolitik*, 16(3):289–310, Sept. 2023.

J. Renshaw and T. Hunnicutt. Biden, Xi agree that humans, not AI, should control nuclear arms. https://www.reuters.com/world/biden-xi-agreed-that-humans-not-ai-should-control-nuclear-weapons-white-house-2024-11-16/, 2024. [Accessed 23-05-2025].

J.-P. Rivera, G. Mukobi, A. Reuel, M. Lamparth, C. Smith, and J. Schneider. Escalation risks from language models in military and diplomatic decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 836–898. ACM, June 2024. doi: 10.1145/3630106.3658942. URL http://dx.doi.org/10.1145/3630106.3658942.

P. Scharre. *Army of none*. WW Norton, New York, NY, Mar. 2019.

U. Securities and E. Commission. AT&T SEC Filing. https://www.sec.gov/ix?doc=/Archives/edgar/data/0000732717/000073271724000046/t-20240506.htm, 2024. [Accessed 23-05-2025].

C. Sienko. The Breach of Anthem Health — The Largest Healthcare Breach in History. https://www.infosecinstitute.com/resources/healthcare-information-security/the-breach-of-anthem-health-the-largest-healthcare-breach-in-history/, 2017. [Accessed 23-05-2025].

Z. Simas. Unpacking the MOVEit Breach: Statistics and Analysis — emsisoft.com. https://www.emsisoft.com/en/blog/44123/unpacking-the-moveit-breach-statistics-and-analysis/, 2023. [Accessed 23-05-2025].

P. W. Singer and A. Friedman. *Cybersecurity and cyberwar*. What Everyone Needs To Know (R). Oxford University Press, New York, NY, Dec. 2013.

R. Sparrow. Killer robots. *J. Appl. Philos.*, 24(1):62–77, Feb. 2007.

B. Stauffer. A Hazard to Human Rights — hrw.org. https://www.hrw.org/report/2025/04/28/hazard-human-rights/autonomous-weapons-systems-and-digital-decision-making, 2025. [Accessed 23-05-2025].

J. D. Steinbruner. Nuclear decapitation. *Foreign Policy*, (45):16, 1981.

UK Ministry of Defence. *Cyber Primer, Third Edition*. UK Ministry of Defence, Development, Concepts and Doctrine Centre, Shrivenham, UK, 2022. URL https://www.gov.uk/mod/dcdc.

United States Air Force. Air force doctrine publication 3-12: Cyberspace operations, February 2023. URL https://www.doctrine.af.mil/Portals/61/documents/AFDP_3-12/3-12-AFDP-CYBERSPACE-OPS.pdf.

J. van Diggelen, E. Aidman, J. Rowa, and J. Vince. Designing ai-enabled countermeasures to cognitive warfare, 2025. URL https://arxiv.org/abs/2504.11486.

L. D. Welch. Cyberspace – the fifth operational domain. *IDA Research Notes*, 2011. URL https://www.ida.org/-/media/feature/publications/2/20/2011-cyberspace---the-fifth-operational-domain/2011-cyberspace---the-fifth-operational-domain.ashx.

A. S. Wilner. US cyber deterrence: Practice guiding theory. *J. Strat. Stud.*, 43(2):245–280, Feb. 2020.

P. Withers. What is the utility of the fifth domain? *Royal Air Force Air Power Review*, 18(1):126–150, 2015.

M. Xiang, R. Fernando, and B. Wang. On-device qwen2.5: Efficient llm inference with model compression and hardware acceleration, 2025. URL https://arxiv.org/abs/2504.17376.

M. Zhang, J. Cao, X. Shen, and Z. Cui. Edgeshard: Efficient llm inference via collaborative edge computing, 2024. URL https://arxiv.org/abs/2405.14371.