Policy Brief

Reducing the Risks of Artificial Intelligence for Military Decision Advantage

Authors
Wyatt Hoffman
Heeu Millie Kim

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

March 2023

Executive Summary

Strategists in the United States and China see the potential of artificial intelligence (AI) to enable better, faster decision-making, which will be decisive in future military conflicts. Applications of machine learning will increasingly influence how political and military leaders perceive the strategic environment, weigh risks and options, and judge their adversaries. But what are the risks of exposing critical human decision-making processes to the surprising behaviors and bizarre failures of AI systems?

Reaping the benefits of AI for decision advantage requires first understanding its limitations and pitfalls. AI systems make predictions based on patterns in data. There is always some chance of unexpected behavior or failure. Existing tools and techniques to try and make AI more robust to failures tend to result in trade-offs in performance, solving one problem but potentially worsening another. There is growing awareness of AI vulnerabilities and flaws, but there is also a need for deeper analysis of the potential consequences of technical failures within realistic deployment contexts.

This policy brief examines how failures in AI systems directly or indirectly influencing decision-making could interact with strategic pressures and human factors to trigger escalation in a crisis or conflict:

- Offensive operations incorporating AI or interfering with an adversary's AI systems could result in unforeseen system failures and cascading effects, triggering *accidental escalation*.
- AI systems that are insecure, inadequately trained, or applied to the wrong types of problems could inject bad information into decision-making processes, leading to *inadvertent escalation*.
- Discovery of a compromise of an AI system could generate uncertainties about the reliability or survivability of critical capabilities, driving decision makers toward *deliberate escalation* if conflict appears imminent.

These scenarios reveal a core dilemma: decision makers want to use AI to reduce uncertainty, especially when it comes to their awareness of the battlefield, knowing their adversaries' intentions and capabilities, or understanding their own capacity to withstand an attack. But by relying on AI, they introduce a new source of uncertainty in the likelihood and consequences of technical failures in AI systems.

Harnessing AI effectively requires balancing these trade-offs in an intentional and risk-informed approach. There is no way to guarantee that a probabilistic AI system will

behave exactly as intended, or that it will give the correct answer. However, militaries can design both AI systems and the decision-making processes reliant upon them to reduce the likelihood and contain the consequences of AI failures, including by:

- Defining a set of mission-specific properties, standards, and requirements for AI systems used in decision-making contexts, such as confidence metrics and safeguards to detect compromise or emergent properties.
- Circumscribing the use of AI in decision-making, applying AI toward narrow questions where it is well suited while reserving for human judgment problems such as interpreting an adversary's intent; and considering ruling out AI in some areas altogether.
- Involving senior decision-makers as much as possible in development, testing, and evaluation processes for systems they will rely upon, as well as educating them on the strengths and flaws of AI so they can identify system failures.

The United States should continue to lead in setting the global standard for responsible development and use of AI by taking steps to demonstrate certain assurances and, to the extent possible, encourage China to take similar precautions by:

- Clarifying the practices and safeguards implemented to limit the risks of AI systems supporting decision-making.
- Collaborating internationally to develop mutually-beneficial technical safeguards and best practices to reduce the risk of catastrophic AI failures.
- Committing to restraint in the use of offensive operations incorporating AI-enabled capabilities and operations targeting AI systems where there are significant escalation risks.

Table of Contents

Introduction

Advanced militaries are pursuing artificial intelligence (AI) to gain decision advantage in future conflicts. Fueling their sense of urgency is the rapid advancement of machine learning applications in recent years toward more general-purpose capabilities that can handle complex, real-world problems. Applications of ML will augment and automate wide-ranging military functions from collecting and fusing intelligence to recommending courses of action to executing cyber and electronic warfare operations. But ML systems do not reason like humans; they look for patterns in data to make predictions. As these systems take on increasingly important roles, how will they impact human decision-making in a crisis or conflict?

This question is salient for three reasons. First, even the best ML systems are often brittle and unreliable. They work well under ideal conditions but quickly fail in the face of unforeseen changes in the environment or malicious interference. There is considerable effort to make *robust* ML systems capable of operating reliably in the face of changing environmental conditions and adversarial interference.[1] But techniques to improve robustness are not mature enough to ensure reliable performance even under relatively benign conditions, much less those of warfare. The need for systems to learn and update in deployment makes this problem even more challenging. As a result, these capabilities could have baked-in vulnerabilities and flaws with potential consequences that are not fully understood.

Second, militaries will face competitive pressures to adopt ML systems despite their uncertain reliability. Militaries are developing ML applications to gain decision advantage in a variety of contexts including intelligence, surveillance, and reconnaissance (ISR), decision support, and cyber and electronic warfare. The fear of falling behind an adversary's AI capabilities may trigger a race to deploy systems with unproven reliability.[2] This temptation may be particularly strong for seemingly innocuous applications in support of human decision-making or in nonlethal roles such as cyber operations.

Third, these powerful but flawed ML capabilities could plausibly be deployed on both sides of a U.S.-China confrontation already fraught with escalation risks. With the bilateral relationship becoming increasingly fractious, commentators have begun to raise the alarm over the increased risk of escalation.[3] However, both countries share a mutual interest in avoiding war. Decision makers on each side are aware of the potential risks of deploying untested and unreliable ML capabilities. Still, the fear of the other side gaining an edge may be stronger than concerns over fielding unproven

systems.[4] The greatest threat may be the mutual risk posed by the deployment of flawed ML systems that undermine stability.

Moreover, ML will become a significant asset to militaries, and could be applied in ways that reduce risks in crisis and conflict. This combination of promise and peril makes it essential to understand the potential impacts of the technical characteristics, quirks, and flaws of these capabilities—especially where they could intersect with strategic concerns like crisis management. This policy brief focuses specifically on risks arising from *how* states design and develop AI systems, and how these systems are incorporated into critical functions that directly or indirectly influence decision-making. It leverages the growing body of technical literature on ML failure modes to consider how these could interact with the strategic pressures and human factors giving rise to escalation risks. The focus on these worst-case scenarios is not to argue that military AI will be inherently destabilizing—like other technologies, AI capabilities could interact with other risk factors in both positive and negative ways.[5] The aim here is to identify the most dangerous possibilities in order to provide recommendations to reduce risks.

This policy brief begins with the technical context: how ML could be applied in the near to medium term to gain decision advantage in warfare, how ML systems can fail in the face of dynamic conditions or adversarial interference, and why this problem is unlikely to be eliminated through technical solutions alone. It then considers the strategic context: how AI might be deployed in a future U.S.-China military confrontation. It then maps potential ML risks onto existing escalation pathways. Finally, it identifies possible measures to mitigate these risks in the development and deployment of ML systems.

The Technical Context: Applying AI in the Military Domain

The relentless advance of ML over the last decade has brought military applications of AI closer to reality.[6] The combination of cheaper computing power, mass data collection, and deep neural network architectures fueled a step change in the capability of ML systems. These systems are not just getting better at a growing range of tasks, they are evolving toward more *general-purpose* capabilities that can solve problems with less reliance on human expertise and narrow, controlled settings.

The relentless advance of ML over the last decade has brought military applications of AI closer to reality.

MuZero, one of DeepMind's game-playing reinforcement learning-based systems announced in 2019, garnered far less attention than its predecessor, AlphaGo.[7] AlphaGo beat world champion Lee Sedol at Go in 2016, revealing the power of reinforcement learning to search for and discover optimal strategies in a highly complex game. But it still relied on a foundation of human knowledge and pre-defined, fixed rules and parameters that made the challenge tractable. MuZero surpassed its performance, but more importantly it did so without any prior knowledge of the game to rely on—just an objective to maximize. Far more so than AlphaGo, it demonstrates a capability that could be applied to real-world problems without predefined rules or set boundaries, where the number of possible moves and counter-moves may be virtually infinite, and where there may be little human expertise to draw from.

Indeed, MuZero served as the basis for an "AI copilot" developed by the U.S. Air Force to locate adversary air defenses. It took only five weeks and five hundred thousand simulations to train "ArtuMu" before giving it control over radar and sensors in a test flight in December 2020.[8] The air force is now developing "CetuMu" with jamming capabilities to thwart ArtuMu in simulated matches; this type of adversarial training is a powerful technique, which may reveal new electronic warfare tactics only feasible for systems operating at machine speed.[9]

For now, real-world ML applications like ArtuMu remain largely experimental. But the progression from AlphaGo to ArtuMu suggests that the significant potential of ML is getting closer to the battlefield. It may soon become technically feasible for militaries to apply ML in a variety of roles for decision advantage—despite whether or not it is wise to do so.

Prospects for ML Applications for Decision Advantage

Decision advantage means making better and faster decisions than an adversary. A military gains it by deploying superior capabilities not only to collect and harness information, but also to protect crucial information and communication flows—and target those of an adversary. ML applications in a range of areas may contribute to decision advantage, many of which are far more mature than experiments at the cutting edge, like ArtuMu.

- **ISR:** ML can automate aspects of collection and processing, such as the identification of objects, selection of potential targets for collection, and guidance of sensors. ML capabilities for image and audio classification and natural language processing are already being used in nonmilitary contexts such as autonomous vehicles and translation services. These use cases have direct parallels in military contexts. ML could also fuse data from multiple sources, dramatically reducing the time from collection to dissemination.

- **Decision support:** ML can augment and assist human decision-making in a number of ways. First, ML systems could enhance situational awareness by creating and updating in real time a common operational picture derived from multiple sensors in multiple domains. Second, ML could perform planning and decision support functions, including matching available weapon systems to targets, generating recommended courses of action and assessing the likelihood of success for various options.[10] These functions are particularly important for coordinating joint operations across domains (space, air, land, sea, and cyber).

- **Electronic warfare:** ML is well suited for tasks such as analyzing, parsing, and filtering signals in support of electromagnetic spectrum operations. These functions have become more technically demanding as radar and countermeasures have evolved. Innovations, such as in infrared search and track, have expanded the range of the electromagnetic spectrum capable of being harnessed. But the more game-changing possibility is "cognitive electronic warfare"—the use of AI to enable capabilities that adapt automatically to adversary tactics and synthesize countermeasures in real time.[11]

- **Cyber warfare:** Cybersecurity is ripe for applications of ML. On the defensive side, ML-enabled intrusion detection can leverage vast amounts of data on network activity to spot anomalous behavior, while ML-enabled antivirus systems can discover patterns in malware to identify unseen or evasive

variants.[12] While more speculative, on the offensive side, attackers might use ML-enabled capabilities to probe adversary networks for weaknesses, gain access and spread through networks more stealthily. ML might assist with specific offensive challenges like developing payloads to manipulate industrial control systems, which can require extensive domain-specific knowledge.[13]

While it is easy to overhype ML, these applications are at least plausible in the near to medium term for three reasons. First, these applications are *technically feasible*, even if some are not necessarily *likely* to reach a level of maturity required for deployment. For instance, ML has been used for years in limited roles such as malware analysis and intrusion detection in cybersecurity, but applications based on reinforcement learning remain experimental.[14] Second, these applications may be more *politically palatable* or less likely to engender the level of resistance that lethal autonomous weapon systems face. Militaries may be more willing to deploy systems with greater degrees of autonomy if they do not trigger as many vexing policy, legal, and ethical questions as LAWS. Finally, these applications are in areas widely considered to be *strategically imperative* for decision advantage in future conflicts, fueling the potential pressures to develop and deploy them. With dwindling barriers to deployment, it is worth exploring the potential implications of these capabilities.

Failure Modes in ML

ML risks fall into two broad categories: safety and security. The former includes the risks of ML systems unintentionally failing due to flaws in the training process or changing conditions in the deployment environment. The latter includes the risks of intentional interference by a malicious actor causing an ML system to fail.[15]

Unintentional failure modes can result from training data that do not represent the full range of conditions or inputs that a system will face in deployment.[16] The environment can change in ways that cause the data used by the model during deployment to differ substantially from the data used to train the model. Such a "distributional shift" in the data often results in rapid degradation in the accuracy of the model's predictions.[17] For instance, ML models used by Amazon were not trained on "pandemic behavior" like bulk-buying products, and their accuracy plummeted at the outset of the COVID-19 pandemic.[18] A flawed training process can also create an ML system that accomplishes its primary task, but does so in a manner contrary to the developer's intent, referred to as specification gaming.[19] An ML system might "hack" its reward, such as the system that won a tic-tac-toe competition by causing opponent systems to crash by requesting nonexistent moves.[20] The system has no intent to cheat at the task, but simply finds the most efficient way to accomplish a predetermined objective.

On the other hand, intentional failure modes include attempts to trick or evade ML models by manipulating the inputs to a system. "Adversarial examples"—inputs (such as images) subtly altered to fool an ML classifier—demonstrate such evasion attacks.[21] Other attacks include directly hacking data feeds and sensors or shaping the deployment environment to deceive a system. In one demonstration, researchers at the Chinese company Tencent placed stickers on a road to trick the lane recognition system in a Tesla self-driving car, causing it to veer into the wrong lane.[22] Malicious actors could also sabotage ML systems during development. Data poisoning attacks corrupt training data to undermine a model's integrity. An attacker with access to a training dataset could simply switch labels on the data to degrade the model's performance. More insidiously, an attacker could create a backdoor so that a model responds to a specific input in a way favorable to the attacker. For example, adding a small number of carefully fabricated samples to a training set for an antivirus system could cause it to misclassify any sample with a specific digital watermark as benign.[23] Adding the watermark to a piece of malware would enable it to slip past defenses undetected.

The Robustness Problem

With growing awareness of these vulnerabilities, AI strategies and proposals increasingly acknowledge the need to make ML *robust* to safety and security risks. Robustness refers to the ability of ML to perform correctly and reliably under conditions that would trigger failure modes, including changes in the environment and adversarial interference.[24] Researchers have proposed many techniques to reduce vulnerabilities and defend ML models against attacks. Yet, there are at least three reasons to be pessimistic about the prospects for achieving this goal in the foreseeable future:

With growing awareness of these vulnerabilities, AI strategies and proposals increasingly acknowledge the need to make ML robust to safety and security risks.

First, attackers continue to beat defenders. A group of researchers put into stark relief the challenge facing ML security when, in 2020, they broke a slate of leading proposed defensive measures against evasion attacks.[25] In many cases, the defenses simply had not been properly evaluated against strong attacks.[26] Defenses against evasion attacks, in the words of two researchers, were "playing a game of whack-a-mole: they close some vulnerabilities, but leave others open."[27] While

prospective measures could prevent certain attacks like poisoning training data, ML systems' fundamental susceptibility to deception will not be resolved anytime soon.

Second, testing and evaluating systems for vulnerabilities can be extremely difficult. ML systems are highly sensitive to changes to input data. Even a subtle alteration to just the right features of an input can completely fool a system. Testing an ML system by presenting it with a wide range of likely inputs cannot guarantee that an attacker will not find one that tricks the system. Testing and evaluation are even more challenging for systems that self-adapt in deployment. For instance, an anomaly detection system for cybersecurity continuously updates its assumptions about normal and anomalous network behavior. This can not only invalidate testing and verification, but also make the system potentially vulnerable to a patient attacker that gradually tweaks normal behavior to avoid appearing anomalous.[28]

Third, fixing ML vulnerabilities often creates other problems.[29] To address vulnerabilities in ML systems, developers have to retrain the system so that it is no longer susceptible to that deception. Retraining is not only expensive, but also has diminishing returns in terms of addressing the underlying issue. For a typical deep learning system, one recent study concluded that "to halve the error rate, you can expect to need more than 500 times the computational resources."[30] Developers may determine that marginal improvements to security are not worth exponentially increasing costs. Retraining may not even be feasible if developers cannot afford to take a system offline or if an attacker has undermined the integrity of training data entirely.

Equally problematic, fixes like retraining a system with synthetically-generated adversarial inputs can negatively impact a system's performance.[31] In one experiment, training a robotic system with adversarial inputs to make it robust against attacks that manipulated sensors inadvertently made it less accurate overall and more prone to accidents.[32] Hardening a system to one set of attacks from a simulated adversary can make it vulnerable to a novel attack.[33] It is possible that training the air force's ArtuMu through repeated competitions against the hostile CetuMu agent might make it more effective, but in the process cause it to develop hidden failure modes that could be triggered by a novel opponent.[34]

At the root of these problems is the fact that ML relies on correlations in data, not understanding causal relationships. This methodology can be incredibly powerful for solving certain problems, but it is also a source of weakness. Existing techniques to improve robustness cannot change this fundamental characteristic. Instead, they tend to make trade-offs, making the system perform better under one set of conditions but

potentially worse in others.[35] This is a problem in adversarial contexts; an attacker can adapt its behavior to exploit the lingering weaknesses of the system.

These persistent problems with safety and security raise the question of whether decision makers will *trust* applications of ML for decision advantage. Operators and military commanders need to trust that ML systems will operate reliably under the realistic conditions of a conflict.[36] Ideally, this will militate against a rush to deploy untested systems. However, developers, testers, policymakers, and commanders within and between countries may have very different risk tolerances and understandings of trust in AI. Moreover, the pressure to avoid falling behind in the deployment of AI may trigger a "race to the bottom" on AI safety, resulting in the fielding of unreliable systems.[37]

The Strategic Context: AI in a U.S.-China Military Confrontation

Harnessing AI has become a priority for both U.S. military and People's Liberation Army (PLA) modernization efforts. Strategists on both sides argue that AI will usher in a new "revolution in military affairs"—a shift from network-centric warfare to decision-centric warfare.[38] Each side is developing capabilities and operational concepts to prepare for this new form of warfare. While still largely aspirational, these efforts offer insights into how each side may deploy AI systems for decision advantage in a potential conflict.

Harnessing AI has become a priority for both U.S. military and People's Liberation Army (PLA) modernization efforts.

AI in PLA Military Strategy and Modernization

Chinese strategists, decision makers, and Xi Jinping himself appear convinced that AI will revolutionize warfare.[39] PLA strategists envision a progression from "informatized" warfare enabled by information and communications technologies to "intelligentized" warfare, which leverages AI, big data, cloud computing and related technologies.[40] Under such conditions, PLA strategists believe wars will be won through "systems destruction warfare" focused on paralyzing and destroying the operational systems of adversary forces.[41] By asymmetrically targeting an adversary's command, control, and communications, the PLA aims to overcome a conventionally superior military.

PLA spending on the development of AI-enabled systems is now upwards of $1.6 billion according to analysis of PLA procurement data by the Center for Security and Emerging Technology (CSET).[42] While much of this spending goes toward the development of autonomous systems and other support functions such as logistics and predictive maintenance, a major area of focus is developing capabilities to prevail in systems destruction warfare. Naturally, a key driver of these investments is the PLA Strategic Support Force (SSF), created in 2015 to consolidate PLA space, cyber, electronic, and information warfare capabilities.[43] Examples of PLA efforts to apply AI for decision advantage include the following:

- **ISR:** The PLA is applying ML to fuse information from systems and sensors across all domains of warfare to improve situational awareness and decision-making.[44] This includes capabilities for fusing satellite data and multisource sensor data, particularly in the maritime domain.[45] The PLA also seeks to

enhance early-warning capabilities through the "intelligentized analysis" of massive data via deep learning.[46]

- **Decision support:** Inspired by AlphaGo's success, China's Central Military Commission Joint Staff Department called for a joint operations command system employing AI to support decision-making in 2016.[47] Around the same time, PLA units began experimentation with such capabilities including an intelligentized joint operations C2 demonstration system developed by the National University of Defense Technology, described as an "external brain" for commanders.[48] Chinese AI companies like DataExa advertise combat decision support services, including the real-time prediction of the movement of foreign weapons platforms.[49] While progress toward operational deployment is difficult to assess, PLA strategists believe AI-enabled command decision-making is not only possible, but inevitable.[50]

- **Electronic warfare:** The PLA is pursuing AI to enable it to manage an ever more dynamic and contested electromagnetic space. PLA procurement contracts include systems for automatic frequency modulation, microwave jamming, broadband automatic gain control, and multisource signal separation; CSET's analysis found numerous contracts focused on "jamming and blinding enemy sensor networks and using AI for cognitive electronic warfare."[51]

- **Cyber warfare:** Available evidence suggests PLA investments into AI-enabled cyber capabilities are focused largely on improving its defenses, such as an SSF contract for an AI-enabled "cyber threat intelligent sensing and early warning platform."[52] A recent report from the Ministry of National Defense claims that the National Defense University's School of Electronic Warfare has developed AI-based capabilities for automated network defense.[53] However, research and experimentation with ML could be utilized offensively. Chinese universities with known connections to state-sponsored hacking groups conduct research on ML security and cyber applications.[54] There is also evidence of experimentation with AI-enabled capabilities at cyber ranges used to practice offensive and defensive cyber operations.[55]

AI in U.S. Military Strategy and Modernization

In 2014, AI became a priority for the U.S. military under the Third Offset Strategy, which sought to harness advanced technologies to offset increases in Chinese and Russian conventional capabilities.[56] Four years later, the Department of Defense (DOD) announced its Artificial Intelligence Strategy and created the Joint Artificial Intelligence

Center to guide the integration of AI systems into decision-making and operations across the military.[57] The JAIC was subsequently incorporated into the Chief Digital and Artificial Intelligence Office.

The JAIC's 2021 AI baseline inventory included 685 AI-related projects and accounts across the DOD.[58] Applications related to decision advantage in warfare likely comprise a small fraction, but include priority areas for modernization. Examples of such efforts include the following:

- **ISR:** AI-enabled collection and fusion of data from multiple domains into a common operational picture is central to the DOD's vision for Joint All-Domain Command and Control (JADC2).[59] JADC2 incorporates a range of initiatives by individual services. For instance, the U.S. Army's Tactical Intelligence Targeting Access Node program aims to use ML to synthesize data from ground, aerial, space, and aerospace sensors.[60] Similarly, the air force is researching the use of algorithms to process and fuse sensor data in its Advanced Battle Management System.[61]

- **Decision support:** The DOD's JADC2 Strategy states that it "will leverage Artificial Intelligence and Machine Learning to help accelerate the commander's decision cycle."[62] In March 2021, Northern Command tested AI-enabled "decision aids" designed to enable domain awareness, information dominance, and cross-command collaboration.[63]

- **Electronic warfare:** A 2016 study by the DOD's Defense Science Board argued that the automation of sensors, communications, and jamming coordination using AI could "achieve information dominance while imposing high 'cost' and disruption on adversaries."[64] The DOD's 2020 Command, Control, and Communications (C3) Modernization Strategy calls for the application of AI to enable "agile electromagnetic spectrum operations" in support of C3.[65] The U.S. military is already incorporating capabilities that facilitate the analysis of signals across the electromagnetic spectrum to better adapt to adversary systems into operational electronic warfare systems.[66]

- **Cyber warfare:** Research applying ML to counter cyberattacks includes the Defense Advanced Research Projects Agency's Harnessing Autonomy for Countering Cyberadversary Systems program, focused on automatically taking down botnets by identifying infected computers, exploiting vulnerabilities to access them, and removing botnet implants.[67] Partnerships under the army's Collaborative Research Alliances include research on AI-enabled cyber

defenses and counter-AI measures.[68] More ambitiously, DOD's Project IKE seeks to create a common command and control architecture for cyber operations capable of generating different courses of action, and even automatically executing operations using AI.[69] The objective, as described by the DOD, is to use AI/ML to "assist human understanding of the cyber battlespace, support development of cyber warfare strategies and measure and model battle damage assessment."[70]

Taking Stock: AI in a U.S.-China Confrontation

Many of the desired capabilities are likely years away from being field-ready. Even after a system is fully trained, tested, and evaluated, it faces numerous hurdles to deployment. Militaries must integrate ML systems into existing legacy systems, doctrine, and operational planning. Human personnel must be trained, operational concepts developed, and organizational structures adapted. In short, it should not be overstated how close AI is to being used on the battlefield.

It should not be overstated how close AI is to being used on the battlefield.

What is increasingly clear, however, is that this is not purely science fiction. Five to ten years out, it is plausible that AI will begin to impact strategic decision-making. ML will shape the information used in critical decision-making processes and influence leaders' perceptions of their adversaries. It could take on more of the burden in developing and evaluating options in operational planning and execution. Tactical-level decisions from ML systems may guide operations in areas where speed and complexity overwhelm human operators. This makes it prudent to try and anticipate and proactively manage the risks these systems could introduce.

AI Failures and Escalation Risks

Numerous studies have cataloged the possible destabilizing effects of AI, including its potential to accelerate the speed of conflict, generate disinformation to sow confusion, or even undermine nuclear deterrence.[71] This report leverages the vast technical literature detailing how and why ML fails to focus on those escalation risks arising from the interaction of technical failures with strategic pressures and human factors in a crisis or conflict. But before examining specific escalation pathways, it is worth noting several features present in the U.S.-China context that make the possible introduction of flawed AI capabilities particularly concerning.

First, the strategic advantage from targeting an adversary's ML capabilities in a future conflict will motivate more aggressive actions to compromise an adversary's ML capabilities *in anticipation of* a conflict. As militaries come to rely on AI systems, we can expect efforts to trigger failure modes in adversary systems to become a regular part of warfare—whether directly (e.g., through hacking) or indirectly, such as by altering tactics to throw off predictions. Acquiring detailed insights into a target ML system, including its architecture, parameters, or training data, enables more reliable and sophisticated attacks.[72] U.S. and PLA strategists alike already recognize that the ability to degrade or destroy an adversary's information and decision-making capabilities will require extensive operational preparation of the environment (OPE) in advance of conflict.[73] The deployment of ML capabilities will likely further incentivize actions prior to a conflict to hunt for vulnerabilities, compromise training processes, or even sabotage capabilities during development.

Second, psychological pressures could compound the impacts of system failures in a crisis. Crises and conflicts put intense pressures on decision makers to respond quickly to developments and act faster than their opponents. Under such conditions, ML system failures may be more likely to go undetected or unremedied. There simply may not be enough time to second guess a system's output, such as a questionable indicator of warning. Decision makers may be more prone to automation bias, or the tendency to accept uncritically the assessments provided by a system.[74] Time pressures can also magnify the impacts of surprise, such as the alarm created by the discovery of a compromised system.[75] If a compromise or failure is detected, there might not be enough time to diagnose the problem, understand its origin or impacts, and take steps to fix it like retraining a system.

Third, preexisting sources of misperception could compound misunderstandings arising from incidents involving ML systems. For example, Chinese strategists appear to overestimate U.S. technological capabilities in general and AI capabilities specifically.[76]

They may be prone to view any behavior or impact of a U.S. operation as intentional, and doubt an explanation that an effect was the result of technical malfunction. Moreover, there is a lack of common understanding of what each side would view as escalatory, particularly in the context of cyber operations.[77] For instance, Chinese decision makers may underestimate how alarming actions such as a cyber intrusion into sensitive ISR systems might be. They may overestimate U.S. detection and diagnosis capabilities and assume that the U.S. could quickly assess an operation and discern the intent behind it.[78] ML systems add another layer of complexity to the challenge of understanding an adversary's capabilities and anticipating how they might perceive or misperceive actions.

Fourth, ML systems may amplify the fears and worst-case assumptions of their makers. Chinese strategists tend to fixate on "false negatives" in early warning, or the risk of failing to detect an incoming U.S. attack.[79] Such fears could influence the creation of training data or simulations in ways that bias an ML model so that it interprets ambiguous actions as positive indicators of an imminent attack. By focusing on minimizing false negatives in the design and training of an ML system, the developers may unintentionally increase the potential for false positives—i.e., misinterpreting incoming data as indication of an attack.

Finally, few crisis management mechanisms exist to provide an off-ramp if an incident triggers escalation. The United States and China have pursued hotlines in various forms, most notably the Defense Telephone Link created in 2008, but it has been rarely used.[80] Amid increasingly frequent Chinese incursions into Taiwanese and Japanese airspace, U.S. officials have expressed concern over the lack of any reliable, direct lines of communication.[81] Efforts to develop stronger military-to-military and leader-to-leader hotlines have run into structural differences between the two militaries, including how each side divides up theater commands.[82] Cultural differences create further impediments. Suspicious of U.S. motives, some Chinese officials fear crisis hotlines would embolden the United States to act more aggressively and view transparency measures as a tool for spying.[83] While U.S. officials view these measures as ways to build trust, for Chinese counterparts, building trust must precede mechanisms for transparency.[84] Such lack of communication and mutual understanding will make it harder for both sides to manage any future incident, let alone one involving unexpected behavior by ML capabilities.

Escalation Pathways

There is little indication that, in their pursuit of AI capabilities for decision advantage, either the United States or China contemplates turning strategic decision-making over

to machines. As long as humans remain in the driver's seat, escalation will largely be a function of the political, strategic, and psychological factors at play during a crisis rather than technology.[85] These factors include uncertainties in decision-making—about an adversary's intentions and resolve, their capabilities, or the consequences of letting them make the first move. If AI systems work as intended, they could reduce these uncertainties, which might be stabilizing. But failures in AI systems could interact with these uncertainties in ways that lead to misperception and miscalculation. It is useful to map these possible interactions onto well-established escalation pathways defined by the underlying cause:[86]

Accidental escalation results when one side's action has unforeseen and unintended impacts that provoke the other side to respond more intensely.

Inadvertent escalation results from an intentional action by one side which triggers an unexpected escalated response. For example, an attack that crosses an undeclared red line.

Deliberate escalation results from one side's actions creating circumstances in which the other views escalation as rational or necessary—though not necessarily based on a completely rational or informed calculation.

Accidental Escalation

ML capabilities could increase the risk of accidentally escalatory impacts of offensive operations in two ways. First, operations targeting an adversary's ML systems might have unpredictable effects. For the same reason that ML systems are vulnerable to begin with—their sensitivity to subtle changes to input data—it can be difficult for an attacker to predict precisely how manipulating those inputs will affect behavior. Realistically, the attacker is unlikely to have complete knowledge of the inner workings of the target system. They will likely have to develop an attack on a substitute model similar to the target model, and hope that the attack will effectively transfer to the actual target. However, in practice, the attack might result in unexpected effects. An operation aiming to degrade an ML system's performance might break it completely.

More widespread and indiscriminate damage can result from operations targeting either publicly available datasets, open-source tools, or shared models.[87] For instance, the effects of poisoning shared datasets could cascade to any model trained on that data, making the overall impact even less predictable. In a similar fashion, corrupting a base model used to train others could result in widespread compromise to systems well beyond the initial target. An adversary might sabotage a development process

without realizing that it could corrupt a system deployed in a critical role, such as for early warning.

Second, the incorporation of ML into offensive capabilities may increase the potential for unexpected and unintended impacts. An ML capability might not be adequately trained for the target environment or could seek to achieve an objective in a way contrary to the operator's intentions. These concerns are particularly acute with cyber operations. Global cyber incidents like NotPetya in 2017 demonstrate the potential for malware to spread far beyond an initial target, with cascading effects.[88] Unintended impacts of a cyber operation could be highly escalatory, particularly if they affect highly sensitive command and control systems—a major concern in the U.S.-China context.[89] Consider a scenario where an ML-enabled cyber operation targets an adversary's space-based assets in an attempt to signal resolve, but rather than temporarily disrupting communications the capability propagates to guidance systems that cause damage to the satellite. The adversary could easily misinterpret the operation as a deliberate strike.

Inadvertent Escalation

ML systems that are insecure, inadequately trained, or applied to the wrong kinds of problems could fail in ways that inject bad information into decision-making, resulting in actions that inadvertently escalate a crisis or conflict. ML vulnerabilities raise obvious concerns, such as the threat from an adversary or a third party hacking a system to create a deception (e.g., making a commercial airplane look like a military one).

Less obvious are the ways in which ML systems could undermine decision-making by simply being applied to problems for which AI is not well suited. As Goldfarb and Lindsay argue, determining whether AI is suitable depends on the type of data, nature of the problem, and the strategic context. They conclude that, "intelligence about 'puzzles' (such as the locations and capabilities of weapon systems) may be more reliable than intelligence about 'mysteries' (such as future intentions and national resolve)."[90]

States in a crisis or at war are unlikely to behave the same as they do during peace. Like the retail algorithms that could not cope with the shifts in consumer behavior in the pandemic, a system trained on years of historical data from peacetime might have its assumptions and expectations upended by a sudden change in the strategic landscape.

Flawed ML-based assessments of an adversary's actions or intent might lead to misperception and miscalculation. Drawing insights from war games, Wong et al. describe one scenario in which an ML system misinterprets an action designed to signal the desire to de-escalate, such as the opponent pulling back forces, as a mistake by the opponent creating an opportunity to gain the upper hand.[91] The system might recommend courses of action that inadvertently escalate the confrontation.

Psychological factors could compound these risks. If a flawed assessment serves to confirm preexisting suspicions regarding an adversary's intentions, it may be less likely to be questioned or scrutinized. Thus, if one side assumes that an adversary is likely to launch a surprise attack, they might not question a system's warning that an attack is likely even if that warning is based on highly ambiguous moves. Even if there is warranted skepticism toward an ML system there might not simply be time to interrogate a system before a response is required.

Deliberate Escalation

ML systems might contribute to a situation in which one side feels pressured to deliberately escalate from fear of suffering an imminent attack or the loss of crucial warfighting capacities. This pressure arises in part from perceptions of significant first-mover advantages from launching a surprise attack, especially with cyber and electronic warfare capabilities, to cripple an opponent's C3 and ISR capabilities. Each side may be prone to assume that at the outset of a conflict the other would attack and try to paralyze those assets, creating strong temptations to strike first and preemptively.[92] Dependence upon vulnerable ML systems could heighten such fears.

In this unstable context, even ambiguous moves by one side could be misinterpreted as preparation for an attack. This is especially true for cyber operations. Offensive cyber operations may require significant activity in advance to gain access and lay the groundwork for an attack, such as compromising dependencies to reach sensitive targets and implanting capabilities to enable future access. Militaries are likely to engage in OPE during peacetime to create offensive options for use in potential contingencies. Yet, it is inherently difficult to distinguish cyber intrusions for espionage from OPE or even an attack already underway. If one side detects an intrusion into sensitive systems in the midst of a crisis, it might assume the worst-case scenario, that the intrusion is preparation for an attack. Buchanan and Cunningham argue that such escalatory risks of cyber espionage and OPE are likely underestimated, particularly on the Chinese side, where there appears to be little to no discussion of the possibility of misinterpretation and escalation from cyber espionage and OPE.[93]

The introduction of ML systems amplifies these risks in two ways. First, cyber intrusions targeting ML systems might appear even more threatening to the target state. An adversary's attempts to probe a deployed ML system might be even harder to interpret. It might be impossible to tell whether an adversary has acquired information enabling them to defeat a system. The precise impacts of a compromise discovered in the development process, such as how tampering with training data may have impacted the system's behavior, could be impossible to rapidly assess. Worst-case assumptions might fill the gaps in decision makers' understanding of the possible impacts on a system and the intent behind them.

Second, there may be no viable options to quickly remediate a compromised ML system, creating a "use it or lose it" dilemma for the target that may pressure them to strike. Not only is the process of "patching" or retraining a system often expensive and time consuming, it is also hard to do so without affecting the system's performance in other ways. Consider an ML-based cyber defense protecting critical C3 systems discovered to contain a backdoor implanted by an adversary during training. The backdoor may have been planted well in advance simply to create options, without the attacker even realizing what systems it would eventually expose. However, the target of the intrusion may fear that its C3 systems are fatally exposed with no way of rapidly fixing the vulnerability. Waiting for the adversary to launch a paralyzing attack might risk having to fight without the advantages of crucial C3 capabilities.

The Dilemma of AI for Decision Advantage

These escalation scenarios reveal a core dilemma of military AI: decision makers want to use AI to reduce uncertainty—to see the battlefield more clearly, understand the adversary's intentions and capabilities, increase confidence in the effectiveness of their own capabilities and ability to anticipate or withstand an attack. But the potential unexpected behaviors or failures of AI systems create another source of uncertainty that can lead to misperception and miscalculation. Employing or targeting AI might make offensive operations less predictable. AI failures might lead to a false sense of certainty about an adversary's moves or intentions. Even the possibility of a hidden failure mode could exacerbate fears about the reliability of potentially compromised systems. Harnessing AI effectively requires balancing this trade-off in risks. There is no way to guarantee that a probabilistic AI system will behave exactly as intended, or that it will give the right answer. Developers and decision makers must be cognizant and intentional about where they introduce and how they manage these uncertainties to avoid catastrophic outcomes.

Risk Mitigation

If done right, ML will become a huge asset to militaries—one that could prove stabilizing if it enhances human decision-making. But managing the potential convergence of technical ML risks and human factors at play in a crisis poses a unique challenge. Solving it is not simply a matter of making better ML systems via more training data or larger models. ML development processes often focus on maximizing a system's accuracy, but not all types of errors are equally concerning.[94] Nor are systems that are highly accurate necessarily risk free (as in the case of an ML-enabled cyber capability that accomplishes its objective in a manner contrary to the commander's intent). Thorough testing and evaluation, validation, and verification (TEVV) practices will be essential but need to be tailored to those specific failure modes that could interact with escalation risks. Further, steps to incorporate AI capabilities into broader systems and processes need to assume the potential for system failures, and contain their impacts where they would be catastrophic. This analysis suggests three general steps to reduce the risks of AI applications for decision advantage:

If done right, ML will become a huge asset to militaries—one that could prove stabilizing if it enhances human decision-making.

First, define a set of mission-specific properties, standards, and requirements for ML systems used for decision advantage. Developers should work with end users, including decision makers, to define necessary characteristics for ML systems in crucial application areas to make them more reliable, predicable, and usable. Depending on the application, desired characteristics may include the following:

- Metrics incorporated into systems to quantify uncertainty in predictions, conveying to decision makers the degree of confidence that a prediction is correct.[95]
- Verifiable robustness properties that can eliminate certain classes of attacks or failure modes to enable deliberate and informed trade-offs between different risks.[96]
- Measures to detect distributional shift, abnormal data or adversarial interference during deployment.[97]

- Methods to discover emergent properties. For example, Kenton et al. propose a method of discovering when the system is guided by certain "agent incentives" that might lead to undesirable behavior.[98]

Second, design decision-making processes to limit the potential consequences of ML failures. Technical measures and TEVV practices can only do so much to reduce the risks of failure. Certain scenarios involving ML system failures simply cannot be tolerated. The solution is to circumscribe where and how ML capabilities are integrated into decision-making and operations.

- Refrain from using ML capabilities in certain high-stakes decision-making contexts. Numerous experts have called attention to the risks of AI in decisions involving nuclear operations in particular.[99] In a working paper submitted to the Tenth Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, the United States, the United Kingdom, and France stated that "Consistent with long-standing policy, we will maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment."[100] This echoed previous formal commitments and statements by senior U.S. officials saying in no uncertain terms that AI will not be used to automate nuclear command and control.[101] While these assurances are welcome, proscribing AI in nuclear command and control only mitigates the most direct threat. Applications influencing decisions indirectly, such as assessments of early warning of a nuclear strike, raise escalation concerns that question whether AI should be used at all.

- Limit reliance on ML assessments based on *types of inferences*. While the broader trend in ML development is toward ever larger and more complex models, applying ML for decision-making will be better served by the opposite instinct: decompose problems, apply AI toward narrow questions ("puzzles") where it is well suited while reserving for human judgment problems such as an interpreting an adversary's intent ("mysteries"), where it is unlikely to prove reliable. This approach requires a careful risk-benefit analysis based on an understanding of the strengths and limitations of ML.

- Build in *redundancy* by employing multiple models relying on different data sources. Diverse types of models and data sources limit the potential impacts of changes in data or adversarial compromise of any single system. Additionally, having back-up models enables fall back options in case of a compromise or failure.[102] These steps would limit the perceived and actual harm from the

discovery of a compromise in a deployed system or a sabotaged training process.

- Restrict the *affordances* of ML systems. Affordances refer to the range of possible actions or outputs available to a system. For instance, an ML capability in an autonomous system with access to weapons and navigation systems has greater affordances than an ML capability only capable of accessing navigation systems. Limiting the affordances of the ML capability might entail segmenting off the weapons systems. By analogy, limiting the affordances of systems involved in decision support might include restricting the possible courses of action a system could recommend to those that do not involve nuclear-capable systems (where deployment requires consideration of an additional level of complicated contextual factors). Restricting the affordances of ML systems may be the only surefire way to prevent certain types of mistakes.

Third, prepare senior decision-makers to be informed users of AI systems. The DOD already strives to achieve appropriate levels of trust through training and education of operators.[103] These practices should extend to senior decision-makers relying upon ML systems.

- Involve decision makers as early as possible in the development and TEVV of the systems upon which they will rely. A lack of communication and engagement between the engineers, developers, data scientists, and end users can lead to problems at the design stage that persist through testing and evaluation processes.[104] Moreover, as Flournoy et al. assert, ML capabilities are embedded within larger systems that include human factors that will affect their performance, and therefore cannot be properly evaluated separate from those larger systems.[105] The only way to fully understand how a system informing decision makers will perform is to test it under realistic conditions, including with the actual humans who will ultimately rely upon it.

- Train senior decision-makers to understand how to interpret and judge the reliability of ML outputs and avoid pitfalls such as automation bias. Senior decision-makers should practice making deliberations involving ML systems under realistic conditions, including scenarios in which the systems fail.

Of course, both sides must address escalation risks to ensure stability in a crisis. In addition to general calls for crisis management mechanisms, experts in both the United States and China have proposed dialogues on the risks of AI to crisis stability.[106] Yet prospects for crisis management or confidence-building measures between the two appear remote. The PLA in particular is reluctant to engage in any dialogue on AI.[107] Still, the United States can take modest steps to provide certain assurances regarding how it will adopt AI and encourage China (and other states) to take similar precautions.

Still, the United States can take modest steps to provide certain assurances regarding how it will adopt AI and encourage China (and other states) to take similar precautions.

First, clarify the practices and safeguards to limit the risks of AI in decision-making. The DOD already leads globally in adopting and making transparent the principles guiding its approach to AI.[108] The United States should explore how to provide further assurances with respect to particularly risky applications of ML:

- Provide details on TEVV practices and standards to the degree possible without conveying information on potential vulnerabilities or weaknesses.[109]
- Identify areas in which decision makers will *not* rely on ML capabilities, for example, if they are deemed unreliable for certain judgements about an adversary's deterrence thresholds or similar types of inferences. This might at least signal to China that these risks are taken seriously.

Second, collaborate internationally to develop mutually-beneficial technical safeguards. Imbrie and Kania, among others, propose international cooperation on AI safety analogous to the United States' offer during the Cold War to share with other countries Permissive Action Links designed to prevent unauthorized launch of nuclear weapons.[110] Technical cooperation faces numerous hurdles, including overcoming mutual suspicions and preventing the transfer of capabilities that would improve an adversary's AI systems. Informal academic or technical-level exchanges are most plausible. Two candidate areas in which to explore mutually-beneficial collaboration include:

- Interpretability and explainability techniques to ensure that decision makers are able to interrogate the outputs of systems.[111]

- Best practices to reduce AI accidents and certain types of failures like reward hacking, which could trigger accidental escalation.[112]

Third, commit to restraint in offensive operations carrying significant escalation risks. It is in the United States' interest to exercise restraint in the conduct of actions carrying significant escalation risks and to encourage such restraint by others. This is true even if it is unlikely that China or others would commit to such norms for the foreseeable future. Two areas stand out as acutely concerning from an escalation perspective:

- Offensive cyber operations: Noting the risks of malware spreading uncontrollably and causing unintended harm, Adams et al. propose that states constrain automation in cyber operations, including via built-in kill switches or incorporating "conditional execution logic" that would prevent an operation from impacting certain targets and restrict effects.[113] The use of ML in offensive cyber operations may compound these automation risks. In such a highly complex and interdependent operational environment, inadequate training or poorly-specified objectives might result in a capability that causes unintended, cascading effects.

- Adversarial AI: Operations directly targeting AI/ML systems used in an adversary's decision-making and C3 may simply be deemed too risky to even attempt, because of the unpredictable ways this could impact decisions. Similarly, operations prior to a conflict aiming to poison or sabotage ML development could have unforeseen, cascading effects.[114]

Conclusion

As they strive for decision advantage, militaries must beware the pitfalls of AI. ML has both the potential to dramatically improve the speed and effectiveness of decision-making and to introduce new uncertainties that could lead to catastrophic miscalculation. The United States leads in establishing policies and processes that will position it to manage these risks.[115] In China, there are at least signs that technical experts are cognizant of the flaws and limitations of AI systems and their potentially destabilizing impacts.

The test will be the resilience of processes and practices against the pressures to fast-track development and deployment to avoid falling behind. A crisis or conflict could rapidly change the risk-benefit calculus for deploying certain applications or pressure militaries to shortcut the development and testing of systems. Moreover, in the absence of clear firebreaks between tactical and strategic-level decision-making, there exists the possibility of a steady creep of AI and automation into influential roles through the incremental shift of more and more parts of analytical and deliberative processes to machines. This makes it all the more necessary to explore proactive steps to limit the risks of AI, even as these capabilities continue to mature and evolve.

Authors

Wyatt Hoffman is a research fellow with the CyberAI Project at CSET. He currently serves as an emerging technology policy fellow in the Office of Emerging Security Challenges at the U.S. Department of State. The views expressed herein are the authors' and do not necessarily reflect those of the U.S. government.

Heeu Millie Kim is a former semester research analyst with the CyberAI Project at CSET.

Acknowledgments

For helpful feedback, suggestions, and critiques, the authors would like to thank Lora Saalman, Haydn Belfield, Shelton Fitch, and John Bansemer.

Document Identifier: doi: 10.51593/2021CA008

Endnotes

[1] For a definition and overview of the concept of robustness, see Elham Tabassi et al., "A Taxonomy and Terminology of Adversarial Machine Learning" (National Institute of Standards in Technology, October 30, 2019), https://doi.org/10.6028/NIST.IR.8269-draft; and Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Robustness and Adversarial Examples" (Center for Security and Emerging Technology, March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/.

[2] Paul Scharre, "Debunking the AI Arms Race Theory," *Texas National Security Review* 4, no. 3 (Summer 2021): 121–32.

[3] See, for instance, Kurt M. Campbell and Ali Wyne, "The Growing Risk of Inadvertent Escalation Between Washington and Beijing," *Lawfare,* August 16, 2020, https://www.lawfareblog.com/growing-risk-inadvertent-escalation-between-washington-and-beijing.

[4] An analysis of articles written by PLA officers and Chinese defense industry engineers and academics on the risks of AI found that, while stability was a major concern, the most frequently cited concern was the potential threat that U.S. AI capabilities might pose to Chinese air defenses, command and control systems, and China's ability to respond to an attack. See Ryan Fedasiuk, "Chinese Perspectives on AI and Future Military Capabilities" (Center for Security and Emerging Technology, August 2020), https://cset.georgetown.edu/publication/chinese-perspectives-on-ai-and-future-military-capabilities/.

[5] On the argument for treating technology as an "intervening variable" rather than a deterministic factor in escalation, see Caitlin Talmadge, "Emerging technology and intra-war escalation risks: Evidence from the Cold War, implications for today," *Journal of Strategic Studies* 42, No. 8 (2019): 864-887.

[6] ML is an approach to, or subfield of, AI. While other AI techniques are relevant to military applications, this study is focused on ML applications, and uses AI and ML interchangeably.

[7] Julian Schrittwieser et al., "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," *Nature* 588, no. 7839 (December 2020): 604–9.

[8] Patrick Tucker, "The Air Force Used AI to Operate the Radar on a U-2 Spy Plane," *Defense One*, December 16, 2020, https://www.defenseone.com/technology/2020/12/air-force-used-ai-operate-radar-u-2-spy-plane/170813/; Will Roper, "The Air Force Flew an AI Copilot on a U-2. Now, the Algorithm Has a New Mission," *Popular Mechanics*, January 19, 2021, https://www.popularmechanics.com/military/research/a35252840/air-force-ai-u2-spy-plane-algorithm-next-mission/.

[9] Will Roper, "The Air Force Flew an AI Copilot on a U-2. Now, the Algorithm Has a New Mission."

[10] Johan Schubert et al., "Artificial Intelligence for Decision Support in Command and Control Systems," ICCRTS 2018: 23rd International Command and Control Research & Technology Symposium, November

2018, https://www.foi.se/download/18.41db20b3168815026e010/1548412090368/Artificial-intelligence-decision_FOI-S--5904--SE.pdf.

[11] Bryan Clark et al., "Winning the Invisible War: Gaining an Enduring U.S. Advantage in the Electromagnetic Spectrum" (Center for Strategic and Budgetary Assessments, 2019), https://csbaonline.org/uploads/documents/Winning_the_Invisible_War_WEB.pdf#page=13; John G. Casey, "Cognitive Electronic Warfare: A Move Towards EMS Maneuver Warfare," *Over the Horizon*, July 3, 2020, https://othjournal.com/2020/07/03/cognitive-electronic-warfare-a-move-towards-ems-maneuver-warfare/.

[12] Micah Musser and Ashton Garriott, "Machine Learning and Cybersecurity: Hype and Reality" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/machine-learning-and-cybersecurity/.

[13] Ben Buchanan et al., "Automating Cyber Attacks: Hype and Reality" (Center for Security and Emerging Technology, November 2020), https://cset.georgetown.edu/publication/automating-cyber-attacks/; Dakota Cary and Daniel Cebul, "Destructive Cyber Operations and Machine Learning" (Center for Security and Emerging Technology, November 2020), https://cset.georgetown.edu/publication/destructive-cyber-operations-and-machine-learning/.

[14] Musser and Garriott, "Machine Learning and Cybersecurity: Hype and Reality."

[15] Ram Shankar Siva Kumar et al., "Failure Modes in Machine Learning Systems," arXiv:1911.11034 [cs.LG] November 25, 2019, http://arxiv.org/abs/1911.11034.

[16] For an overview of issues associated with ML training data in a military context see Arthur Holland Michel, "Known Unknowns: Data Issues and Military Autonomous Systems" (United Nations Institute for Disarmament Research, May 17, 2021), https://unidir.org/known-unknowns.

[17] Joaquin Quiñonero-Candela et al., "Dataset Shift in Machine Learning," MIT, https://mitpress.mit.edu/books/dataset-shift-machine-learning.

[18] Will Douglas Heaven, "Our Weird Behavior during the Pandemic Is Messing with AI Models," *MIT Technology Review*, May 11, 2020, https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/.

[19] Rudner and Toner, "Key Concepts in AI Safety: Specification in Machine Learning."

[20] For a survey of similar cases see Joel Lehman et al., "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities," *Artificial Life* 26, no. 2 (April 9, 2020): 274–306.

[21] Ian J. Goodfellow et al., "Explaining and Harnessing Adversarial Examples," arXiv:1412.6572 [stat.ML], December 20, 2014, http://arxiv.org/abs/1412.6572.

[22] Evan Ackerman, "Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve Into Wrong Lane," *IEEE Spectrum*, April 1, 2019, https://spectrum.ieee.org/three-small-stickers-on-road-can-steer-tesla-autopilot-into-oncoming-lane.

[23] Giorgio Severi et al., "Exploring Backdoor Poisoning Attacks Against Malware Classifiers," arXiv:2003.01031 [cs.CR], March 2, 2020, http://arxiv.org/abs/2003.01031.

[24] Tabassi et al., "A Taxonomy and Terminology of Adversarial Machine Learning."

[25] Tramer et al. tested and defeated 13 techniques employed to defend image classifiers against adversarial examples. Florian Tramer et al., "On Adaptive Attacks to Adversarial Example Defenses," arXiv:2002.08347 [cs.LG], February 19, 2020, http://arxiv.org/abs/2002.08347.

[26] Nicholas Carlini, "Are Adversarial Example Defenses Improving?" February 20, 2020, https://nicholas.carlini.com/writing/2020/are-adversarial-exampe-defenses-improving.html.

[27] Ian Goodfellow and Nicolas Papernot, "Is Attacking Machine Learning Easier than Defending It?" cleverhans blog, February 15, 2017, http://www.cleverhans.io/security/privacy/ml/2017/02/15/why-attacking-machine-learning-is-easier-than-defending-it.html.

[28] Wyatt Hoffman, "Making AI Work for Cyber Defense: The Accuracy-Robustness Tradeoff" (Center for Security and Emerging Technology, December 2021), https://cset.georgetown.edu/publication/making-ai-work-for-cyber-defense/.

[29] Andrew J. Lohn and Wyatt Hoffman, "Securing AI: How Traditional Vulnerability Disclosure Must Adapt" (Center for Security and Emerging Technology, March 2022), https://cset.georgetown.edu/wp-content/uploads/Securing-AI.pdf.

[30] Neil C. Thompson et al., "Deep Learning's Diminishing Returns," *IEEE Spectrum*, September 24, 2021, https://spectrum.ieee.org/deep-learning-computational-cost.

[31] Dimitris Tsipras et al., "Robustness May Be at Odds with Accuracy," arXiv:1805.12152 [stat.ML], May 30, 2018, http://arxiv.org/abs/1805.12152.

[32] Mathias Lechner et al., "Adversarial Training Is Not Ready for Robot Learning," arXiv:2103.08187 [cs.LG], March 15, 2021, http://arxiv.org/abs/2103.08187.

[33] Florian Tramèr and Dan Boneh, "Adversarial Training and Robustness for Multiple Perturbations," arXiv:1904.13000, 2019, https://arxiv.org/abs/1904.13000.

[34] Gleave et al. find that "even apparently strong self-play policies can harbor serious but hard to find failure modes." Adam Gleave et al., "Adversarial Policies: Attacking Deep Reinforcement Learning," arXiv:1905.10615 [cs.LG], May 25, 2019, http://arxiv.org/abs/1905.10615.

[35] Wyatt Hoffman, "Making AI Work for Cyber Defense: The Accuracy-Robustness Tradeoff."

[36] Heather M. Roff and David Danks, "'Trust but Verify': The Difficulty of Trusting Autonomous Weapons Systems," *Journal of Military Ethics* 17, no. 1 (2018): 2–20.

[37] Paul Scharre, "Debunking the AI Arms Race Theory."

[38] See, for instance, Bryan Clark et al., "Mosaic Warfare: Exploiting Artificial Intelligence and Autonomous Systems to Implement Decision-Centric Operations" (Center for Strategic and Budgetary Assessments, 2020), https://csbaonline.org/uploads/documents/Mosaic_Warfare.pdf; Elsa B. Kania, "Artificial intelligence in China's revolution in military affairs," *Journal of Strategic Studies* 44, no. 4 (2021).

[39] Elsa B. Kania, "Chinese Military Innovation in Artificial Intelligence," Testimony before the U.S.-China Economic and Security Review Commission Hearing on Trade, Technology, and Military-Civil Fusion, June 7, 2019, https://s3.us-east-1.amazonaws.com/files.cnas.org/backgrounds/documents/June-7-Hearing_Panel-1_Elsa-Kania_Chinese-Military-Innovation-in-Artificial-Intelligence.pdf?mtime=20190617115242&focal=none.

[40] Ibid.

[41] Jeffrey Engstrom, *Systems Confrontation and System Destruction Warfare: How the Chinese People's Liberation Army Seeks to Wage Modern Warfare* (RAND Corporation, 2018), https://www.rand.org/pubs/research_reports/RR1708.html.

[42] Ryan Fedasiuk et al., "Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence" (Center for Security and Emerging Technology, October 2021), https://cset.georgetown.edu/publication/harnessed-lightning/.

[43] Elsa B. Kania and John Costello, "Seizing the Commanding Heights: The PLA Strategic Support Force in Chinese Military Power," *Journal of Strategic Studies* 44, no. 2 (2021): 218–64; Ryan Fedasiuk et al., "Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence."

[44] Elsa B. Kania, "Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power" (Center for New American Security, November 28, 2017), https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-military-revolution-and-chinas-future-military-power.

[45] Ryan Fedasiuk et al. "Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence."

[46] Elsa B. Kania, "Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power."

[47] Ibid.

[48] Elsa B. Kania, "Chapter 20: Artificial Intelligence in Future Chinese Command Decision-Making," in Nicholas D. Wright (ed.) *AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives* (NSI Inc., 2018), https://nsiteam.com/social/wp-content/uploads/2019/01/AI-China-Russia-Global-WP_FINAL_forcopying_Edited-EDITED.pdf.

[49] Ryan Fedasiuk et al., "Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence."

[50] Elsa B. Kania, "Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power."

[51] Ryan Fedasiuk et al. "Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence."

[52] Ibid.

[53] CNA, "The China AI and Autonomy Report," Issue 2, November 9, 2021, https://www.cna.org/Newsletters/China-AI/The-China-AI-and-Autonomy-Report-Issue-2.pdf.

[54] Dakota Cary, "Academics, AI, and APTs: How Six Advanced Persistent Threat-Connected Chinese Universities Are Advancing AI Research" (Center for Security and Emerging Technology, March 2021), https://cset.georgetown.edu/publication/academics-ai-and-apts/.

[55] Dakota Cary, "Downrange: A Survey of China's Cyber Ranges" (Center for Security and Emerging Technology, September 2022), https://cset.georgetown.edu/publication/downrange-a-survey-of-chinas-cyber-ranges/.

[56] Gian Gentile et al., *A History of the Third Offset, 2014-2018* (RAND Corporation, 2021), https://www.rand.org/pubs/research_reports/RRA454-1.html.

[57] Department of Defense, "Summary of the 2018 Department of Defense Artificial Intelligence Strategy," https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

[58] Government Accountability Office, "Artificial Intelligence: DOD Should Improve Strategies, Inventory Process, and Collaboration Guidance," March 2022, www.gao.gov/assets/gao-22-105834.pdf.

[59] Sherrill Lingel et al., "Joint All-Domain Command and Control for Modern Warfare" (RAND Corporation, 2020), https://www.rand.org/content/dam/rand/pubs/research_reports/RR4400/RR4408z1/RAND_RR4408z1.pdf.

[60] Courtney Albon, "Army developing TITAN system to improve sensor-to-shooter timeline," *Defense News,* October 6, 2022, https://www.defensenews.com/land/2022/10/06/army-developing-titan-system-to-improve-sensor-to-shooter-timeline/.

[61] Margarita Konaev et al., "U.S. Military Investments in Autonomy and AI: A Budgetary Assessment" (Center for Security and Emerging Technology, October 2020), https://cset.georgetown.edu/publication/u-s-military-investments-in-autonomy-and-ai-a-budgetary-assessment/.

[62] U.S. Department of Defense, "Summary of the Joint All-Domain Command & Control (JADC2) Strategy," March 2022, https://media.defense.gov/2022/Mar/17/2002958406/-1/-1/1/SUMMARY-OF-THE-JOINT-ALL-DOMAIN-COMMAND-AND-CONTROL-STRATEGY.PDF.

[63] Theresa Hitchens, "Exclusive: NORTHCOM Developing, Testing AI Tools To Implement JADC2," *Breaking Defense*, March 5, 2021, https://breakingdefense.com/2021/03/exclusive-northcom-developing-testing-ai-tools-to-implement-jadc2/.

[64] Defense Science Board, "Report of the Defense Science Board Summer Study on Autonomy," Office of the Under Secretary of Defense, June 2016.

[65] Department of Defense, "Command, Control, and Communications Modernization Strategy," September 2020, https://dodcio.defense.gov/Portals/0/Documents/DoD-C3-Strategy.pdf.

[66] Bryan Clark et al. "Winning the Invisible War."

[67] Defense Advanced Research Projects Agency, "Harnessing Autonomy for Countering Cyberadversary Systems," https://www.darpa.mil/program/harnessing-autonomy-for-countering-cyberadversary-systems.

[68] Konaev et al., "U.S. Military Investments in Autonomy and AI: A Budgetary Assessment."

[69] Zachary Fryer-Biggs, "Twilight of the Human Hacker" (the Center for Public Integrity, September 13, 2020), https://publicintegrity.org/national-security/future-of-warfare/scary-fast/twilight-of-the-human-hacker-cyberwarfare/.

[70] Mark Pomerleau, "Cyberwarriors Lack Planning Tools. That Could Change," *Fifth Domain*, November 25, 2019, https://www.fifthdomain.com/dod/2019/11/25/cyber-warriors-lack-planning-tools-that-could-change/.

[71] See, inter alia, Rebecca Hersman, "Wormhole Escalation in the New Nuclear Age," *Texas National Security Review,* Summer 2020, https://tnsr.org/2020/07/wormhole-escalation-in-the-new-nuclear-age/; Michael C. Horowitz, "When speed kills: Lethal autonomous weapon systems, deterrence and stability," *Journal of Strategic Studies* 42, no. 6 (2019); Michael C. Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures" (Center for New American Security, January 2021), https://s3.us-east-1.amazonaws.com/files.cnas.org/backgrounds/documents/AI-and-

International-Stability-Risks-and-Confidence-Building-Measures.pdf?mtime=20210112103229&focal=none; James S. Johnson, "Artificial Intelligence: A Threat to Strategic Stability," *Strategic Studies Quarterly*, Spring 2020.

[72] Wyatt Hoffman, "AI and the Future of Cyber Competition" (Center for Security and Emerging Technology, January 2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/.

[73] Ben Buchanan and Fiona S. Cunningham, "Preparing the Cyber Battlefield: Assessing a Novel Escalation Risk in a Sino-American Crisis," *Texas National Security Review* 3, no. 4 (2020), https://tnsr.org/2020/10/preparing-the-cyber-battlefield-assessing-a-novel-escalation-risk-in-a-sino-american-crisis/.

[74] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York, NY: W.W. Norton & Company, 2018), 144.

[75] Ariel E. Levite et al., "China-U.S. Cyber-Nuclear C3 Stability" (Carnegie Endowment for International Peace, April 2021), https://carnegieendowment.org/files/Levite_et_all_C3_Stability.pdf.

[76] Fedasiuk, "Chinese Perspectives on AI and Future Military Capabilities."

[77] Jason Healey and Robert Jervis, "The Escalation Inversion and Other Oddities of Situational Cyber Stability," *Texas National Security Review* 3, no. 4 (Fall 2020): 30–53.

[78] Levite et al., "China-U.S. Cyber-Nuclear C3 Stability."

[79] Lora Saalman, "Fear of False Negatives: AI and China's Nuclear Posture," *Bulletin of the Atomic Scientists*, April 24, 2018, https://thebulletin.org/2018/04/fear-of-false-negatives-ai-and-chinas-nuclear-posture/.

[80] Phelim Kine, "'Spiral into Crisis': The U.S.-China Military Hotline Is Dangerously Broken," *Politico*, September 1, 2021, https://www.politico.com/news/2021/09/01/us-china-military-hotline-508140.

[81] Ibid.

[82] Jack Detsch, "Biden Looks for Defense Hotline With China," *Foreign Policy*, May 10, 2021, https://foreignpolicy.com/2021/05/10/biden-china-xi-jinping-defense-hotline-pentagon/.

[83] Ibid.; Levite et al., "China-U.S. Cyber-Nuclear C3 Stability."

[84] Levite et al., "China-U.S. Cyber-Nuclear C3 Stability."

[85] Talmadge, "Emerging technology and intra-war escalation risks: Evidence from the Cold War, implications for today."

[86] For an overview see "Chapter Two: The Nature of Escalation," in Forrest E. Morgan et al., *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, CA: RAND Corporation, 2008).

[87] On the vulnerabilities of open-source data and tools see Andrew Lohn, "Poison in the Well: Securing the Shared Resources of Machine Learning" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/poison-in-the-well/.

[88] Andy Greenberg, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History," *Wired*, August 22, 2018, https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/.

[89] Levite et al., "China-U.S. Cyber-Nuclear C3 Stability."

[90] Avi Goldfarb and Jon R. Lindsay, "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War," *International Security* 46, no. 3 (2022), https://direct.mit.edu/isec/article/46/3/7/109668/Prediction-and-Judgment-Why-Artificial.

[91] Wong et al. note that "autonomous systems, programmed to take advantage of tactical and operational advantages as soon as they can identify them, might create inadvertent escalation in situations where the adversary could be trying to prevent further conflict and escalation." Yuna Huh Wong et al., "Deterrence in the Age of Thinking Machines" (RAND Corporation, 2020), https://www.rand.org/pubs/research_reports/RR2797.html.

[92] James Johnson, "China's Vision of the Future Network-Centric Battlefield: Cyber, Space and Electromagnetic Asymmetric Challenges to the United States," *Comparative Strategy* 37, no. 5 (2018): 373–90.

[93] Ben Buchanan and Fiona S. Cunningham, "Preparing the Cyber Battlefield: Assessing a Novel Escalation Risk in a Sino-American Crisis."

[94] Violet Turri et al., "Measuring AI Systems Beyond Accuracy," arXiv:2204.04211 [cs.SE], April 7, 2022, https://arxiv.org/pdf/2204.04211.pdf.

[95] Hollen Barmer et al., "Robust and Secure AI" (Carnegie Mellon University, 2021), https://resources.sei.cmu.edu/asset_files/WhitePaper/2021_019_001_735346.pdf.

[96] Yizheng Chen et al., "Learning Security Classifiers with Verified Global Robustness Properties," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)* (November 15–19, 2021), https://arxiv.org/abs/2105.11363.

[97] Tegjyot Singh Sethi et al., "A Dynamic-Adversarial Mining Approach to the Security of Machine Learning," arXiv:1803.09162v1 [cs.LG], March 24, 2018, https://arxiv.org/pdf/1803.09162.pdf.

[98] Zachary Kenton et al., "Discovering Agents," arXiv:2208.08345v2 [cs.AI], August 24, 2022, https://arxiv.org/pdf/2208.08345.pdf.

[99] Michael C. Horowitz and Paul Scharre, "AI and International Security: Risks and Confidence-Building Measures" (Center for New American Security, January 2021), https://s3.us-east-1.amazonaws.com/files.cnas.org/backgrounds/documents/AI-and-International-Stability-Risks-and-Confidence-Building-Measures.pdf?mtime=20210112103229&focal=none; and James Johnson, "AI, Autonomy, and the Risk of Nuclear War," War on the Rocks, July 29, 2022, https://warontherocks.com/2022/07/ai-autonomy-and-the-risk-of-nuclear-war/.

[100] "Principles and responsible practices for Nuclear Weapon States," July 29, 2022, Available at: https://www.un.org/en/conferences/npt2020/documents.

[101] See comments by Lt. Gen. (ret.) Jack Shanahan, former director of the JAIC. Sydney J. Freedberg Jr, "No AI For Nuclear Command & Control: JAIC's Shanahan," Breaking Defense, September 25, 2019, https://breakingdefense.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/. Similar commitments were included in the unclassified version of the 2022 Nuclear Posture Review, https://s3.amazonaws.com/uploads.fas.org/2022/10/27113658/2022-Nuclear-Posture-Review.pdf, and the UK "Defence Artificial Intelligence Strategy," https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy.

[102] Tegjyot Singh Sethi et al., "A Dynamic-Adversarial Mining Approach to the Security of Machine Learning."

[103] DOD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway," June 2022, https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

[104] Hollen Barmer et al., "Robust and Secure AI."

[105] Michele A. Flournoy et al., "Build Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning systems" (WestExec Advisors, 2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

[106] Centre for Humanitarian Dialogue, "Code of Conduct on Artificial Intelligence in Military Systems," August 18, 2021, https://hdcentre.org/insights/ai-code-of-conduct/.

[107] Gregory C. Allen, "One Key Challenge for Diplomacy on AI: China's Military Does Not Want to Talk" (Center for Strategic and International Studies, May 2022), https://www.csis.org/analysis/one-key-challenge-diplomacy-ai-chinas-military-does-not-want-talk.

[108] DOD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway," June 2022, https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

[109] Further study is needed to determine what information could feasibly be shared without revealing vulnerabilities. Moreover, it is doubtful that the PLA would reciprocate in sharing information on its internal practices; but it is worth noting that this recommendation was discussed in the track II dialogue conducted by the Center for Humanitarian Dialogue, which included Chinese participants. See Centre for Humanitarian Dialogue, "Code of Conduct on Artificial Intelligence in Military Systems," August 18, 2021, https://hdcentre.org/insights/ai-code-of-conduct/.

[110] Andrew Imbrie and Elsa B. Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement" (Center for Security and Emerging Technology, December 2019), https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Safety-Security-and-Stability-Among-the-Great-Powers.pdf.

[111] Centre for Humanitarian Dialogue, "Code of Conduct on Artificial Intelligence in Military Systems."

[112] Zachary Arnold and Helen Toner, "AI Accidents: An Emerging Threat" (Center for Security and Emerging Technology, July 2021), https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/.

[113] Perri Adams et al., "Responsible Cyber Offense," *Lawfare*, August 2, 2021, https://www.lawfareblog.com/responsible-cyber-offense.

[114] Wyatt Hoffman, "AI and the Future of Cyber Competition."

[115] These policies include DOD's AI Ethical Principles, Responsible Artificial Intelligence Strategy and Implementation Pathway, and DOD Directive 3000.09.