ARTIFICIAL INTELLIGENCE, REAL RISKS: UNDERSTANDING—AND MITIGATING—VULNERABILITIES IN THE MILITARY USE OF AI

Nick Starck, David Bierbrauer and Paul Maxwell |
01.18.22

Editor's note: This article is part of the Army Cyber Institute's contribution to the series, "Compete and Win: Envisioning a Competitive Strategy for the Twenty-First Century." The series endeavors to present expert commentary on diverse issues surrounding US competitive strategy and irregular warfare

with peer and near-peer competitors in the physical, cyber, and information spaces. The series is part of the Competition in Cyberspace Project (C2P), a joint initiative by the Army Cyber Institute and the Modern War Institute. Read all articles in the series here.

Special thanks to series editors Capt. Maggie Smith, PhD, C2P director, and Dr. Barnett S. Koven.

No one likes to wake up in the morning, but now that artificial intelligence–powered algorithms set our alarms, manage the temperature settings in our homes, and select playlists to match our moods, snooze buttons are used less and less. AI safety-assist systems make our vehicles safer and AI algorithms optimize police patrols to make the neighborhoods we drive through, and live in, safer as well. All around us, AI is there, powering the tools and devices that shape our environment, augment and assist us in our daily routines, and nudge us to make choices about what to eat, wear, and purchase— with and without our consent. However, AI is also there when our smart devices start deciding who among us is suspicious, when a marginalized community is disproportionately targeted for police patrols, and when a self-driving car kills a jaywalker.

AI is becoming ubiquitous in daily life, and war is no exception to the trend. Reporting even suggests that the November 2020 assassination of the top Iranian nuclear scientist was carried out by an autonomous, AI-augmented rifle capable of firing up to six hundred rounds per minute. Russia and China are rapidly developing, and in some cases deploying, AI-enabled irregular warfare capabilities and it is only a matter of time before the same fissures, biases, and undesirable outcomes that are occurring with the AI systems that power our daily lives begin appearing in the AI systems used to wage war and designed to kill.

MOST POPULAR POSTS

Underground Nightmare: Hamas Tunnels and the Wicked Problem Facing the IDF

The Five Reasons Wars Happen

It's Time to Ukrainify US

Given the role of AI and machine learning in strategic competition, it is critical that we understand both the risks introduced by these systems and their ability to create a strategic advantage. By exploring adversarial methods, it is possible to begin building such an understanding. An examination of four categories of adversarial methods offers a window into the vulnerabilities in these systems. In this article, we will use the problem of target identification as the base example and explore how an AI system's learning and thinking can be attacked. Although this example occurs in conflict, the methods we describe can also be used during competition. This analysis leads to two important conclusions: First, a human must remain in the loop in any use of artificial intelligence. And second, AI may not provide a strategic advantage for the United States in the era of great power competition, but we must continue to invest and encourage the ethical use of AI.

Adversarial Methods

Like other military systems, AI systems go through multiple distinct lifecycle phases—development (data collection and training), testing, operation, and maintenance. In each of these phases there are unique vulnerabilities that must be identified and for which we much account. We will proceed through the development of a hypothetical AI target identification system that is learning to identify enemy armored vehicles. At each stage we will explore the associated class of adversarial methods—poisoning, evasion, reverse engineering, and inference—as well as how we might protect our systems from each.

Poisoning

The first step in the development of any AI system is problem identification and data collection. With our challenge to identify enemy armored vehicles, we must define our problem. Do we want to identify all enemy armored vehicles, or only a certain type from a specific adversary? This problem definition informs the collection

Military Assistance

UPCOMING EVENTS

There are no upcoming events.

ANNOUNCEMENTS

Announcing the Modern War Institute...

Essay Contest Call for Submissions: Solving...

Call for Applications: MWI's 2023–24...

Join Us This Friday for a Livestream with...

and preparation of a set of related data, which in this case would include a large number of images of the enemy armored vehicles of interest. Not only must we accumulate images of all the vehicles of interest, but we also need images in a variety of conditions—varying light, differing angles, limited exposure, and alternate channels (e.g., infrared, daylight), for example. The data is then prepared by data analysts for use in the training of the AI system. However, the sheer amount of data required to develop an AI system creates a vulnerability. The volume of data means that analysts do not have the capacity to verify that each collected image is an actual enemy armored vehicle or that the images represent the full range of types of armored vehicles.

This stage of development is the first point at which an adversary can attack the AI system through a technique called poisoning. The goal of poisoning is to alter the data the AI system uses in training so that what the AI learns is flawed. This process attacks the integrity of the system before it ever goes into operation.

The underlying methodology of crafting malicious raw data to induce a flawed analytical outcome is the same as in traditional military deception. Operation Quicksilver, a deception operation prior to the Allied invasion of Normandy during World War II, sought to attack the German defensive analytic model. To accomplish this attack, the Allies created a ghost army (poisoned data) led by Lt. Gen. George Patton to skew the Germans' analysis (the model) of where they should focus their defenses (model output).

Such large-scale deception operations may be more difficult to achieve in today's interconnected society, but poisoning data is feasible. Our adversaries know that we are pursuing AI-enabled target identification. Knowing that such an AI system would need training images of their current armored vehicles, an adversary could poison those training images by manipulating their vehicles' appearance. This could be as simple as the addition of a distinctive symbol like a red star on vehicles that they

suspect may be under surveillance. Our AI system would then be trained on these poisoned images of deliberately manipulated vehicles and "learn" that all enemy armored vehicles have red stars.

Though such a poisoning attack would occur during a state of competition, the impact would manifest in conflict when the adversary deploys armored vehicles without red stars to avoid detection. Further, the adversary could paint red stars on civilian vehicles to induce our AI system to falsely identify the civilians as combatants.

Ensuring our systems learn correctly can be accomplished in many ways. Detailed data curation could help alleviate risk but would consume valuable time and resources. Rather, a scalable solution includes data governance policies to improve the integrity and representativeness of the data used for AI systems. The proper placement of technical controls and well-trained personnel remaining in the loop during all phases of the AI lifecycle will further reduce the risk of a poisoning attack.

Evasion

The next type of attack, evasion, relies on similar fundamental attack principles but deploys them when the AI system is in operation. Instead of poisoning what the AI is learning, an evasion attack targets how the AI's learning is applied. This may sound like a trivial difference; however, it has significant implications on the resources an attacker needs to be successful and conversely what actions a defender needs to take. In the poisoning attack, the attacker needs the ability to control or manipulate the data used to train the model. In an evasion attack, the attacker needs, at a minimum, the ability to control the inputs to the AI system during operation.

Evasion attacks are well suited to computer vision applications such as facial recognition, object detection, and target recognition. A common evasion technique involves slightly modifying the color of certain image

pixels to attack how the system applies what it has learned. To the human eye, it may appear as if nothing changed; however, the AI may now misclassify the image. The effects of this technique were demonstrated by researchers when an AI that previously correctly identified an image of a panda was shown what looked to be the same image but had been manipulated with added colors throughout the image that are imperceptible to the human eye. The AI not only misidentified the panda as a gibbon but did so with remarkably high confidence.

An attacker that also gains access to the system's outputs or predictions could develop a more robust (all images of pandas are misidentified) or targeted (all pandas are seen as another specific animal) evasion method.

The evasion attack principles can also be employed in the physical world—for example, wearing specially made sunglasses to obscure or alter your image on a facial recognition camera. This is the same principle behind camouflage. In this case the adversary is targeting the model's perception rather than a human's. In a military context, if an adversary knew that our AI targeting system was trained on tanks with desert camouflage, the adversary's tanks could simply be repainted in woodland camouflage to deliberately evade detection by our AI systems. An AI-enhanced autonomous scout system may now be unable to effectively recognize targets and fail to provide commanders with timely and accurate intelligence.

Evasion attacks are some of the most widely researched adversarial methods, so defending against all possible attack vectors will prove challenging. Steps to harden our AI systems, however, can increase our overall confidence that they function as intended. One such step would be implementing evaluation tools prior to deployment. These tools test the AI system against a variety of known adversarial methods to give us a quantitative measure of its robustness. Maintaining a human in the loop where possible during operation can also mitigate against evasion attacks.

Reverse Engineering

The previous two classes of attacks shared similar fundamental principles for targeting AI systems during development and operation. These attacks also had natural analogs to traditional military concepts like deception and camouflage. However, the risks to AI systems are not all so straightforward and potential vulnerabilities exist outside of development and operation. What are the vulnerabilities in AI systems while they are in maintenance or storage? What are the risks if an adversary gains access to an AI system through a network intrusion or by capturing a next-generation, AI-enabled drone on the battlefield?

In the class of attacks known as reverse engineering, an adversary attacks an AI system with the goal of extracting what the AI system has learned and, ultimately, enable the model to be reconstructed. To conduct a reverse engineering attack, an adversary needs to be able to send inputs to a model and to observe the outputs. This attack bypasses any encryption or obfuscation of the model itself. For our hypothetical target identification AI, this attack could be conducted by an adversary sending out vehicles of different types (the inputs) and observing which elicit a response from the AI (the outputs). While such an attack would take time and risk the loss of resources, eventually an adversary would be able to learn what the target identification model considered to be a threat.

With this information, the adversary would be able to develop its own version of our AI system. In addition to making other adversarial attacks easier to develop, direct knowledge of how the AI is making its decisions enables an adversary to predict our responses or avoid them entirely. Such an insight into our AI-enhanced decision-making processes would pose a significant threat to our operational security across the conflict continuum.

Protecting our systems against reverse engineering can prove difficult, especially because mission requirements

may require the system to allow for many queries or weighted outputs as opposed to simple binary decisions. This highlights a need for a range of tailored policies to manage the risks associated with adversarial methods. These may include strict accountability of AI-enabled systems, especially those deployed at the edge like drones or smart goggles. Further, we could impose access limitations by only allowing authorized users to view system outputs.

Inference Attacks

The final class of attack, known as inference attacks, is related to reverse engineering. Rather than trying to recover what the AI system learned, an adversary is trying to extract what data the AI system used in its learning process. This is a subtle but meaningful distinction that has significant implications for models trained on sensitive or classified data.

To conduct an inference attack, as with reverse engineering, the adversary needs the ability to send inputs to a model and to observe the outputs. With a set of inputs and outputs, the adversary can train an adversarial AI that predicts if a given data point was used to train our friendly model.

Imagine our target identification AI is trained on classified images of an adversary's new weapons system. Using an inference attack, the adversary could learn that the secrecy of this weapon system had been compromised. In other words, an inference attack on our AI systems could facilitate the compromise of classified intelligence. If this is done during competition, it can have big implications for crisis and conflict.

Much like reverse engineering, managing risk associated with inference attacks will mostly be handled through policy decisions. In addition to the access policy decisions, there will be difficult decisions about when to use sensitive or classified date in the training of AI systems, what type of data to use, and in what quantity. These

decisions will need to balance performance against risks to develop AI systems that can still meet mission requirements.

Implications for Great Power Competition

Of course, this is clearly not an exhaustive explanation of the full range of adversarial methods. However, this framework should provide a sufficient overview with which leaders can explore the full implications, both positive and negative, of the integration of AI systems into our formations. Both the United States and our adversaries are pursuing this technology for an asymmetric advantage in the strategic competition to come, and both sides cannot win such an advantage.

Data Asymmetry

When we think about technology and asymmetric advantage, it is useful to begin with first principles and consider relative access to the "raw" materials. In AI systems, the raw materials are data—vast amounts of data. Does the United States have access to the same quality and quantity of data as our adversaries? Given the legal factors and social norms around privacy and data security in national security in the United States—critical topics in their own right—the answer is not obviously "yes." This suggests that the United States would be at an inherent disadvantage in the development and deployment of AI systems.

Development Capacity

Well-trained personnel are the other critical resource for AI systems. As the Army has identified with its "People First" strategy, having the right personnel will be critical to the United States' success in strategic competition. The United States has talent in industry, academia, and the military. Whether these personnel can be recruited, retained, and directed toward hard national security problems is an open question that is worthy of dedicated thought. In the short term, the talented individuals that

are already within our formations should be identified and the disparate efforts across organizations working on AI should be synchronized.

AI is Only a Tool

Artificial Intelligence is a tool. Like any other tool, it has inherent strengths and weaknesses. Through a deliberate and realistic evaluation of those strengths and weaknesses, the United States can find the optimal balance between the risks and rewards of AI. While AI may not deliver the maximum asymmetric advantage the United States is looking for in strategic competition, neither can we cede the technology to our adversaries who are investing heavily in the field. Instead, the United States can, and should, support the ethical use of AI, promote research into robust AI, and develop defensive best practices for AI systems. Implementing these actions and others, based on an understanding of the vulnerabilities and limitations of AI systems, will lead the United States to more effectively situate artificial intelligence into a strategy for the era of great power competition.

Captain Nick Starck is a US Army cyber officer currently assigned as a research scientist at the Army Cyber Institute. His research focuses on information warfare and data privacy.

Captain David Bierbrauer is a signal officer in the US Army. He earned a master of science in engineering degree for applied mathematics and statistics from the Johns Hopkins University in 2021. Captain Bierbrauer is currently a data engineer and data scientist at the Army Cyber Institute.

Dr. Paul Maxwell was commissioned as an armor officer in 1992 and served as a battalion XO/S-3, brigade S-4, company commander, scout platoon leader, company XO, and mechanized infantry platoon leader. At the United States Military Academy, he has served as an instructor, assistant

professor, and associate professor in the Department of Electrical Engineering and Computer Science. His current position is the deputy director at the Army Cyber Institute at West Point.

The views expressed are those of the authors and do not reflect the official position of the United States Military Academy, Department of the Army, or Department of Defense.

Image credit: Naval Information Warfare Center Pacific (adapted by MWI)