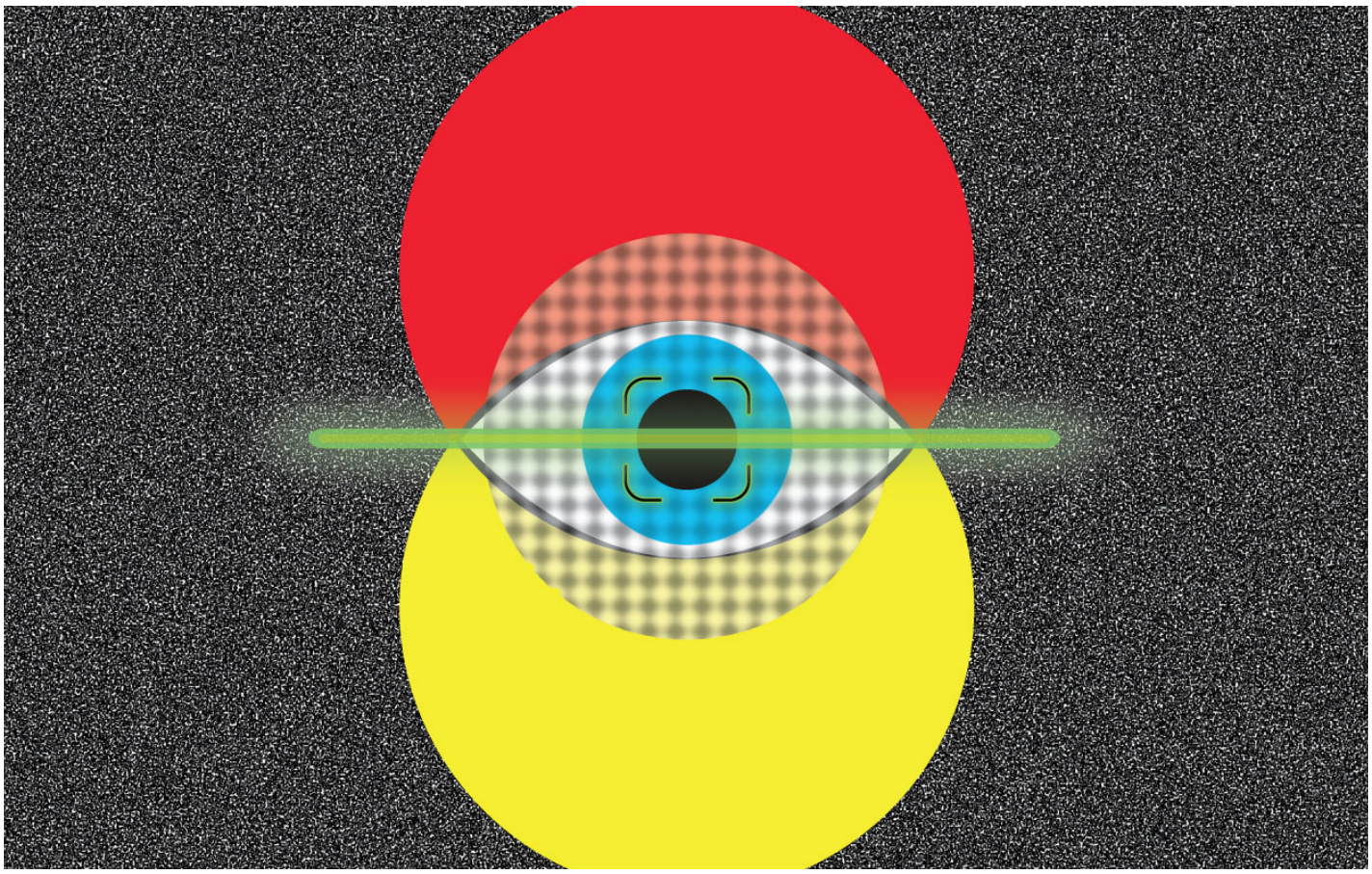
Deloitte.
Insights



FEATURE

Securing government against adversarial AI

Artificial intelligence is bringing new benefits to government, but new adversarial attacks against those tools mean government also needs new ways to safeguard them.

Vinita Fordham, Dr. David Caswell, and Allie Diehl

THE DELOITTE AI INSTITUTE FOR GOVERNMENT AND DELOITTE CENTER FOR GOVERNMENT INSIGHTS

IN 2016, MICROSOFT released its AI-enabled chatbot Tay publicly to social media and retrained it based on inputs from its conversations with users.[1] Shortly after the release, internet trolls launched a coordinated data-poisoning attack that abused Tay's learning mechanisms, enabling the attackers to retrain it to tweet inappropriate content. This AI system's integrity was compromised by bad actors, through *data poisoning*—introducing malicious inputs to manipulate the model's outputs. Machine learning (ML) models trained on open-source data or production data can be especially vulnerable to this type of attack.[2]

As AI/ML solutions proliferate, the attacks on such systems also multiply. Some real-world examples include cybersecurity breaches, privacy attacks on patient records, and intellectual property theft. Attacks can even occur by way of physical manipulation, such as in the case of autonomous vehicles: A research study demonstrated that an autonomous driving AI system could be fooled into driving over the speed limit and misclassifying fake traffic signs.[3]

Since this early attack, Microsoft has published open-source adversarial AI defenses through its publicly available repository, Counterfit.[4] Such precautions are critical to avert successful attacks as more AI-enabled chatbots are released into production, especially within government services. Fortunately, data and technology leaders recognize the need for safeguards. In Deloitte's *2019 State of AI in the Enterprise* report, organizations highlighted security and safety as their top concerns when adopting AI. But this field of knowledge—known as adversarial AI—has only recently emerged.[5] At a practical level, organizations need AI guardrails against adversarial attacks and plans for defensive countermeasures. Addressing these new security challenges requires a multipronged approach, including cross-training of AI/ML and cybersecurity teams, setting security standards and bringing in experts, and securing the development life cycle.

The problem: A new breed of security challenges

As ML applications become more prevalent and applied to areas where reliability, security, and safety are paramount, attention has turned to their potential vulnerabilities. As they upgrade their systems to include AI-based technologies, federal agencies are starting to encounter more adversarial AI challenges. For example, the Department of Justice is addressing data privacy or guarding against data tampering that impact tagging of facial recognition for potential use in court cases, where it would be used to identify suspects.[6]

However, building AI that can defend against this new breed of security challenges is not easy. A small number of government and industry leaders are working toward establishing AI-/ML-specific standards. For example, Defense Advanced Research Projects Agency's (DARPA) Guaranteeing AI Robustness Against Deception (GARD) program is currently working to develop general defensive capabilities that are effective against a broad swath of adversarial attacks. As part of the countermeasures, they have developed an approach to capture improved metrics for AI robustness.[7] These metrics seek to allow for future advances in the field to be shared utilizing a common understand of how they work.

Such examples remain rare, and attention to the challenge of adversarial AI is limited, as the field itself reached critical milestones as recently as six years ago with nearly half its publications occurring since 2020.[8] As a result, government agencies are still working to comply with the 2019 Executive Order on Maintaining American Leadership on AI (EO 13859)[9] to develop "technical standards [to] minimize vulnerability to attacks from malicious actors and reflect federal priorities for innovation, public trust, and public confidence in systems that use AI technologies." The broader ecosystem of AI service providers is also still developing the necessary capabilities. Globally, venture capital

FIGURE 1

AI/ML development is a cyclical process involving a training phase during which iterative development occurs and an inference phase during which insights are drawn from real-world data

Source: Deloitte analysis.

funds invested more than US$230 billion into AI/ML companies from 2018 to 2021, including almost US$90 billion in 2021. However, less than 1% of this funding has gone toward AI/ML security research and infrastructure development startups.[10]

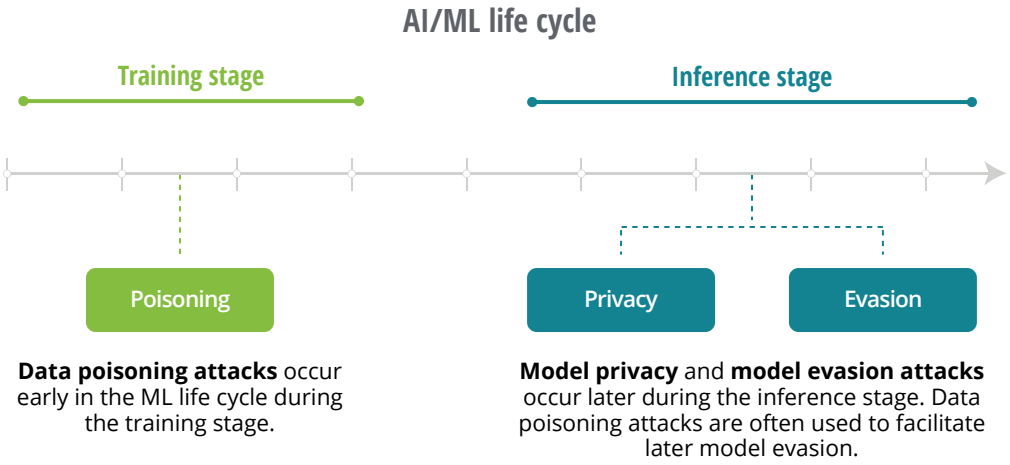
New technology brings new vulnerabilities

One critical shortcoming inhibiting progress in the field of adversarial AI is the lack of a standard framework for technical evaluations and governance. The lack of standardization makes it hard to benchmark the security levels of AI systems. Because this is an emerging field, there is no clear starting point for addressing AI security systematically.

Attacks can take place at any point throughout the AI/ML life cycle (figure 1), from training to inference—and point solutions cannot secure entire AI/ML systems (figure 2). Since defenses geared toward production models will not always account for vulnerabilities in their training, security experts should examine AI/ML systems throughout the entire AI/ML life cycle. For example, encryption techniques that protect cloud-based development activities will not secure a model trained on poisoned data—at that point, it is already too late. What's needed is a mechanism to

FIGURE 2

Types of adversarial AI attacks



Source: Deloitte analysis.

protect AI systems against threats such as data poisoning and associated model evasion—such as through embedded backdoor attacks or attacks which allow for future evasion attacks by creating an opening or "backdoor" for a specific type of adversarial input to trick the model—and privacy attacks in which malicious actors attempt to steal private data such as patient details, information on the data set or model, or the model itself.

Adversarial AI is not just traditional software development

There are marked differences between adversarial AI and traditional software development and cybersecurity frameworks. Often, vulnerabilities in ML models are connected back to data poisoning and other types of data-based attacks. Since these vulnerabilities are inherent in the model inputs, they cannot be "patched" through code fixes the way traditional software fixes are done. Furthermore, methods to patch known vulnerabilities tend to be expensive and only provide modest security improvements while impacting model performance severely. Complicating the problem even further, as models become increasingly bespoke or personalized to users, a single patch is less likely to fix the issue. In fact, it may even exacerbate it. Some adversarial AI countermeasures, while protecting against one type of attack, may increase vulnerability to others. As Andrew Lohn and Wyatt Hoffman, research fellows at Georgetown's Center for Security and Emerging Technology, wrote, "One defense may protect against attacks that alter images to hide a tank, but that same defense may make it easier for attacks to adjust the images in other ways to make tanks appear where there are none."[11]

The solution: A three-pronged approach

Adversarial AI attacks can pose a significant threat to machine learning models, making it important to address these challenges proactively. It is recommended to address these as early as possible in the AI development process and continuously evaluate and update the defenses as the threat

landscape evolves. To address the challenges posed by adversarial AI, organizations should focus on three main areas:

1. **Cross-train the workforce** to bridge the gap between AI/ML and cybersecurity expertise—the intersection of these disciplines provides the best defense against adversarial attacks.

2. **Set security standards and bring in the specialists**, such as through AI red teaming and model governance, to evaluate the security of AI/ML models and suggest countermeasures and risk mitigations.

3. **Secure the model development life cycle** by adopting the tools, techniques, and standards being developed in the rapidly evolving ecosystem around adversarial AI and adopting the most relevant and trusted tools and frameworks.

CROSS-TRAINING THE WORKFORCE

Consider people first when evaluating security risks. Vulnerability to adversarial AI threats starts with the workforce. Those entrenched in the ML life cycle—data scientists, engineers, and other AI talent personas—prioritize model performance, which often carries a trade-off with model security. Organizations can empower their AI teams to better navigate performance-security trade-offs by cross-training the workforce to enable model governance. Data science and cybersecurity are the two technology areas for which organizations have the most difficulty finding talent with the appropriate skillsets.[12] Cross-train AI engineers in security leading practices while cross-training security analysts in AI fundamentals is a great way to effectively identify the threat landscape and attack perimeter. This people-centric approach aligns with trends in the cybersecurity industry, which has moved from the concept of "shifting security to the left" (that is, introducing security checks during the development phase) to enhancing security everywhere. This change has been accelerated by the transition into the virtual world at the onset of the COVID-19 pandemic. As more work is done from the home office, security considerations must be taken across new networks.

Those familiar with the software development fields will recognize the term DevOps, a set of practices that combines software development and IT operations. Similar to DevOps, MLOps is a culture and a practice that aims to combine ML development and operations throughout the model development life cycle, allowing for the continuous iteration and delivery of models. It helps track model development and postproduction performance, and to plan for optimal retraining and redeployment. Successful MLOps can decrease time to production, reduce model downtime, and enhance model performance. By integrating security considerations into an MLOps workflow, organizations can take a proactive approach to combating adversarial AI threats.

There are many roles involved in the ML life cycle, from data engineering to deployment and mission specialists. For an MLOps plus security—or MLSecOps—approach to be successful, the different roles/personas must all work together to ensure model security at every step of the ML life cycle.[13] Defenses cannot be siloed, nor are they the responsibility of a single stakeholder (figure 3). In this way, there is no singular "security persona," but rather a combined effort enabled by experience-based cross-training. Collaboration tools and governance workflows can support coordination across the various personas and ensure that development decisions consider security principles throughout the ML life cycle.

MLSecOps responsibilities span across multiple roles involved in the AI/ML development life cycle. As all ML begins with data, **data engineers** play an important role in ensuring security through data governance. In the first phase of the life cycle, data engineers validate the integrity of training data. This validation step is especially critical given

FIGURE 3

Security through MLSecOps is the responsibility of many different personas involved in the AI/ML life cycle

Role	Training phase	Inference phase
Data engineer Source and prepare datasets for modeling	Validate integrity of training data	Detect/classify malicious input
Data scientist Experiment, apply, tune, and train AI/ML models	Design secure models (utilize privacy-enhancing development techniques e.g., differential privacy, adversarial, training)	Continuously iterate on model design (employ model explainability techniques)
ML engineer Deploy AI models for end user usage	Ensure security of hardware supply chain (e.g., GPUs) and employ encryption techniques	Employ robust logging/tracking capabilities
Business analyst Evaluate model for business alignment	Monitor externally published information	Consider human-in-the-loop integration to validate model outputs

Source: Deloitte analysis.

training data is often pulled from open-source or third-party data providers.[14] During the later inference phase, data engineers employ detection and classification techniques to identify potential adversarial inputs and protect model integrity. **Data scientists** build security into the models by design. As part of the development process, they employ privacy-enhancing techniques such as differential privacy, a mathematical method for ensuring individual data points cannot be extracted from a data set. The role of data scientists also extends into the inference phase as they examine model outputs to detect vulnerabilities. For example, a data scientist may examine a model explainability plot, or a visual representation of model performance, to better understand the model's underlying mechanisms and thus potential security vulnerabilities regarding how it handles edge cases. **ML engineers** further secure the life cycle through use of encryption techniques to protect data integrity throughout as well as validating the security of software and hardware components used during both the training and inference phases. **Business analysts** also have a role to play by setting strong policies around data and model protection, ensuring sourcing hardware from preapproved manufacturers. Further, to minimize leakage of sensitive data and proprietary ML models, they monitor externally published information to protect against reconnaissance efforts intend on informing model attacks. Throughout the life cycle, analysts consider vulnerabilities within the process that could be addressed through technical or nontechnical solutions and apply risk-quantification and benchmarking practices.

SET SECURITY STANDARDS AND BRING IN THE SPECIALISTS

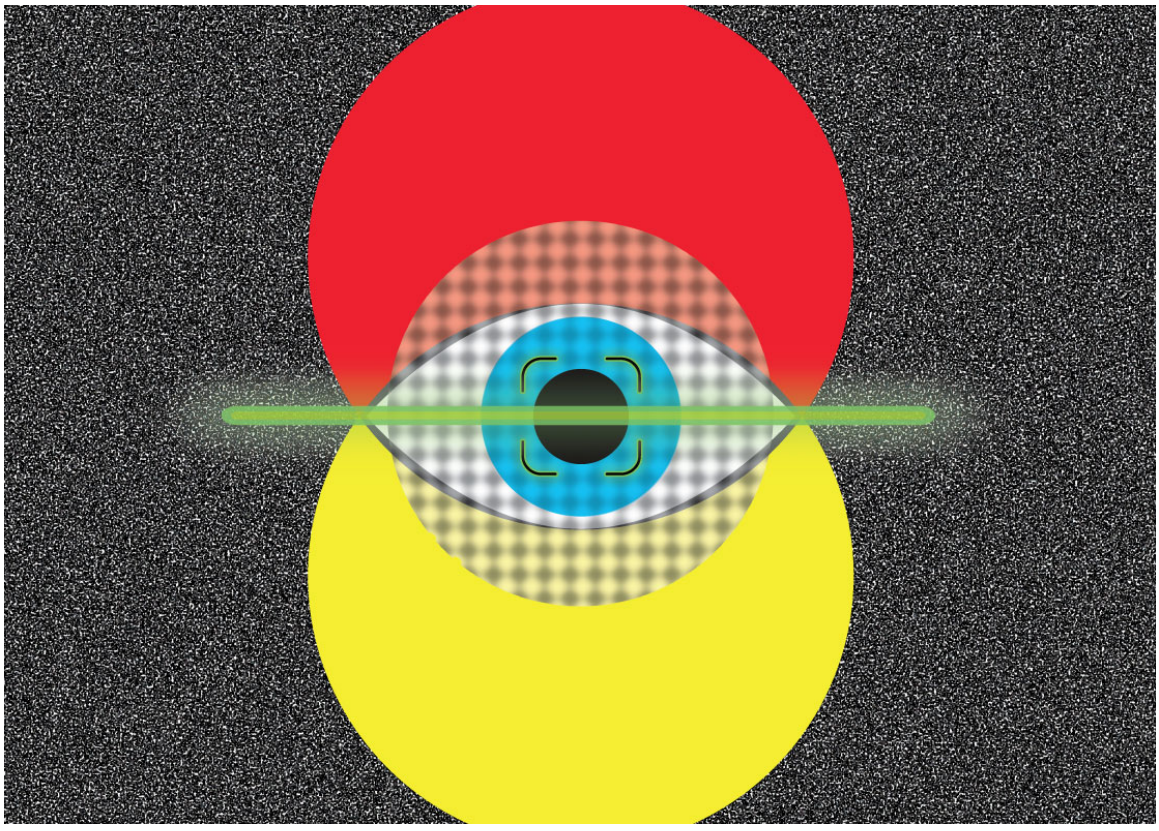
As part of model governance, organizations should develop and maintain a counter–adversarial AI framework. Dependent on the type of ML systems and applications an organization uses, different levels of defense and privacy thresholds may be warranted. Organizations can start with an assessment of AI projects, identifying those that involve critical systems and data and base recommendations on quantitative evidence.

Currently, much of the evidence available about AI threats is based on academic metrics that are not always relevant in actual operations. Organizations should invent ways to measure both the vulnerability and the potential impact of adversaries attacking real-world systems.

Chief security officers, chief data officers, and other C-suite executives focused on AI should be proactive to ensure their organizations are using leading technologies in a rapidly evolving field. This includes understanding and evaluating the ecosystem for adversarial AI solutions when budgeting for investments. Although a small portion of funding is going toward AI/ML security research and infrastructure development, many private companies, academia, and governments are developing and investing in solutions to combat adversarial AI attacks. Academia and big tech lead the way in maintaining open-source libraries as well as through the publication of threat databases.

For example, in academia, researchers at Johns Hopkins University developed a set of tools called TrojAI to help test the robustness of ML models and protect them from trojan attacks. These tools generate data sets and models with trojans to test ML models quickly and at scale.[15] In MIT's Computer Science and Artificial Intelligence Laboratory, researchers developed a tool called TextFooler that uses adversarial text to test robustness of natural language models.[16] Other popular open source frameworks—such as the Adversarial Threat Landscape for AI Systems (ATLAS), developed by MITRE with support from Microsoft and a broad coalition of private sector companies—use real-world examples to catalog possible attacks and corresponding defenses across multiple industry verticals.[17]

As the landscape of both adversarial attacks and defenses is rapidly evolving, organizations should invest in AI red teaming. Originating in the



cybersecurity field, a "red team" (or "ethical hacker") works to expose vulnerabilities with the objective of making a system stronger. These teams should operate independently to assure an objective review as a red team must be able to keep sensitive data secure, while being transparent about identifying risks. "Red teaming" is new to AI but can be traced back to stress testing exercises completed by the Department of Defense back in the 1960s largely for cybersecurity. Now, many large tech companies employ AI red teams to look for AI security vulnerabilities (figure 4).

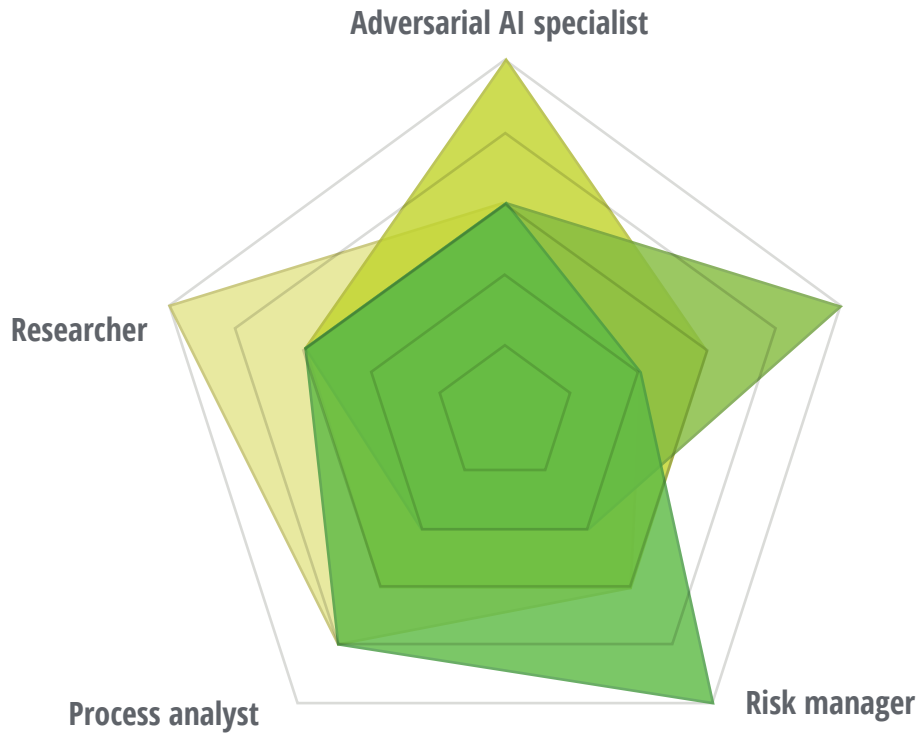
The composition of a red team will depend on the needs of the organization. It may consist of software developers and data scientists or those more focused on human factors or insider threats. If an organization uses multiple models trained on datasets with the same distribution (often the case for large computer vision or language models), an adversarial AI specialist would be able to identify vulnerabilities to highly transferable attacks. However, with unique tabular data (rows and columns), launching a successful attack would require more brute force trial-and-error and would not require an expert.

SECURE THE MODEL DEVELOPMENT LIFE CYCLE

Historically, organizations have focused security budgets on protecting hardware and software against attacks. However, AI models themselves are susceptible to threats outside of the traditional hardware-software dichotomy. Traditional testing methods can assess model robustness against random inputs but do not account for the

FIGURE 4

Composition of an AI red team



Source: Deloitte analysis.

complexity of a targeted attack like data poisoning. One such technique, adversarial training, is a defense technique by which a model is retrained with "adversarial examples" (such as those used in a data poisoning) included in the training data set but with correct labels. This teaches the model to ignore noise and learn from unperturbed features. Adversarial training is, in fact, one of the most popular defense techniques, showing great promise in academic studies.[18]

Yet, like many adversarial AI defenses, it also has its disadvantages—specifically, it can only be used to protect against known attacks. The value and weakness of each defense approach illustrates the importance of exploring adversarial AI across the ML life cycle and applying multiple defenses throughout. While some measures, such as adversarial training, can be utilized by AI/ML developers throughout model development, other defense techniques, such as input detection, can identify unexpected threats later. These can be especially useful in situations where the development team utilizes pretrained/prepackaged models and does not have access to the training phase of the model life cycle.

Now is the time to start

Deploying AI applications, such as intelligent chatbots, into the real world can open the door to adversarial attacks. Combating these risks requires a joint effort by the leaders of data, technology, security, and talent teams. Risk-mitigating decisions should align with the AI/ML life cycle through an MLSecOps approach—for example, validating which training data is being used to retrain a chatbot. Furthermore, governance controls can be put into place to prevent attacks on privacy or thwart hackers attempting to access safeguarded private information.

With the expansions of open-source libraries and tools, fueling the so-called "democratization of AI," there has also been a *democratization of adversarial attacks,* which are becoming increasingly automated. As AI algorithms are increasingly incorporated into business and mission processes, organizations need to monitor the developments and investments in this space to stay up to date on current AI security and make strategic investments to help protect their models.

Endnotes

1. Jing Cao, “Microsoft takes AI bot ‘Tay’ offline after offensive remarks,” Bloomberg, March 24, 2016.
2. Tasha Austin, et al., *Trustworthy open data for trustworthy AI: Opportunities and risks of using open data for AI*, Deloitte Insights, December 10, 2021.
3. Chawin Sitawarin et al., *DARTS: Deceiving autonomous cars with toxic signs*, Princeton University and Purdue University, accessed March 15, 2023.
4. Github, “Azure/counterfit,” accessed March 3, 2023.
5. The National Institute of Technology and Standards (NIST) defines Adversarial AI as “concerned with the design of ML algorithms that can resist security challenges, the study of the capabilities of attackers, and the understanding of attack consequences; see: Elham Tabassi et al., *A taxonomy and terminology of attacks and mitigations of adversarial machine learning,* accessed March 3, 2023.
6. US Government Accountability Office, *Facial recognition technology: Current and planned uses by federal agencies*, August 21, 2023.
7. GARD Project, “Holistic evaluation of adversarial defenses,” accessed March 3, 2023.
8. Eugene Neelou and Alex Polyakov, “The history of adversarial AI,” Hack in the Box Security Conference, accessed March 15, 2023.
9. Federal Register, “Maintaining American leadership in artificial intelligence,” February 14, 2019.
10. Caitlin Pintavorn, “Adversarial attacks and the future of secure AI,” *Medium*, April 8, 2020.
11. Andrew Lohn and Wyatt Hoffman, “How traditional vulnerability disclosure must adapt,” CSET, March 14, 2023.
12. Jen DuBois, “The data scientist shortage in 2020,” Quanthub, 2020; Deloitte, “Solving the cyber talent gap,” January 28, 2021.
13. In his landmark 2016 paper, *Leading adversarial AI,* researcher Nicholas Carlini defines “robustness” as a neural network’s (type of AI/ML) ability to “reduce the success rate of current attacks’ ability to find adversarial examples.” See: Nicholas Carlini and David Wagner, *Towards evaluating the robustness of neural networks,* University of California, Berkely, March 22, 2017
14. Austin, et al., *Trustworthy open data for trustworthy AI*.
15. Johns Hopkins Applied Physical Laboratory, “APL and the intelligence community tackle malware in the age of AI,” press release, August 28, 2020.
16. GitHub, ”Jind11/TextFooler,” accessed March 14, 2023.
17. MITRE, “MITRE ATLAS™,” accessed March 3, 2023.
18. For example, see Uri Shaham, Yutaro Yamada, and Sahand Negahban, *Understanding adversarial training: Increasing local stability of neural nets through robust optimization,* Yale University, January 16, 2016; also see, Aleksander Madry, et al., “Towards deep learning models resistant to adversarial attacks,” Cornell University, September 4, 2019.

Acknowledgments

The authors would like to thank **Josh Rachford**, **Joe Mariani**, and **Pankaj Kishnani** for their support in the development of the draft. The authors would also like thank **Tasha Austin**, **Mekala Ravichandran**, and **Bruce Chew** for providing their insights and thoughtful feedback on the draft.

About the authors

Vinita Fordham | vfordham@deloitte.com

Vinita Fordham is a leader at Deloitte focused on emerging and disruptive technologies, specializing in artificial intelligence in the government sector. Her career has been focused on driving technological paradigm shifts including leading many initiatives applying new technologies to solve unique "business" challenges for government agencies as well as mid- and large-tier companies. Before joining Deloitte, Fordham was a senior executive in the US government, focusing on next-generation technologies and leading large-scale digital transformation efforts.

Dr. David Caswell | dcaswell@deloitte.com

Dr. David Caswell is a managing director in Deloitte Risk & Financial Advisory working within the Cyber portfolio across defense, security, and justice (DS&J) accounts. Dr. Caswell has more than 20 years of experience as a technologist specializing in artificial intelligence (AI) and cybersecurity. He focuses on leveraging AI and analytics to enable cybersecurity for the defense, security, and justice sector. His PhD research from Stanford University's management science and engineering program developed AI technique applications for national policy decisions.

Allie Diehl | aldiehl@deloitte.com

Allie Diehl is a senior consultant in Deloitte's Analytics & Cognitive practice focused on growing AI strategy and adoption across the federal government. She has lectured on AI at Georgetown University and helped design an advanced degree class. She graduated Phi Beta Kappa from The Johns Hopkins University with a double major in economics and international studies as a Hodson Scholar. She has previously worked as an equities research analyst and an actuarial analyst.

Contact us

Our insights can help you take advantage of change. If you're looking for fresh ideas to address your challenges, we should talk.

Industry leadership

Tasha Austin
Principal | Deloitte Risk and Financial Advisory
+1 571 882 5479 | laustin@deloitte.com

Tasha Austin is a principal in Deloitte's Risk and Financial Advisory business and has more than 22 years of professional services experience involving commercial and federal financial statement audits, fraud, dispute analysis and investigations, artificial intelligence, and advanced data analytics.

The Deloitte Center for Government Insights

Joe Mariani
Senior research manager | Deloitte Center for Government Insights
+1 410 576 7618 | jmariani@deloitte.com

Joe Mariani leads the emerging technology research program for Deloitte's Center for Government Insights. His research focuses on the intersection of culture and innovation in both commercial businesses and government organizations.

About the Deloitte AI Institute for Government

The Deloitte AI Institute for Government is a hub of innovative perspectives, groundbreaking research, and immersive experiences focused on artificial intelligence (AI) and its related technologies for the government audience. Through publications, events, and workshops, our goal is to help government use AI ethically to deliver better services, improve operations, and facilitate economic growth.

We aren't solely conducting research; we're solving problems, keeping explainable and ethical AI at the forefront and the human experience at the core of our mission. We live in the Age of With: humans with machines, data with actions, decisions with confidence. The impact of AI on government and its workforce has only just begun.

About the Deloitte Center for Government Insights

The Deloitte Center for Government Insights shares inspiring stories of government innovation, looking at what's behind the adoption of new technologies and management practices. We produce cutting-edge research that guides public officials without burying them in jargon and minutiae, crystalizing essential insights in an easy-to-absorb format. Through research, forums, and immersive workshops, our goal is to provide public officials, policy professionals, and members of the media with fresh insights that advance an understanding of what is possible in government transformation.

About the Deloitte AI Institute

The Deloitte AI Institute helps organizations connect all the different dimensions of the robust, highly dynamic, and rapidly evolving AI ecosystem. The AI Institute leads conversations on applied AI innovation across industries, using cutting-edge insights to promote human-machine collaboration in the Age of With™. The Deloitte AI Institute aims to promote dialogue about and development of artificial intelligence, stimulate innovation, and examine challenges to AI implementation and ways to address them. The AI Institute collaborates with an ecosystem composed of academic research groups, startups, entrepreneurs, innovators, mature AI product leaders, and AI visionaries to explore key areas of artificial intelligence including risks, policies, ethics, future of work and talent, and applied AI use cases. Combined with Deloitte's deep knowledge and experience in artificial intelligence applications, the institute helps make sense of this complex ecosystem and, as a result, delivers impactful perspectives to help organizations succeed by making informed AI decisions. No matter what stage of the AI journey you're in—whether you're a board member or C-suite leader driving strategy for your organization, or a hands-on data scientist bringing an AI strategy to life—the Deloitte AI Institute can help you learn more about how enterprises across the world are leveraging AI for a competitive advantage. Visit us at the Deloitte AI Institute for the full body of our work, subscribe to our podcasts and newsletter, and join us at our meetups and live events—let's explore the future of AI together. Learn more.

Deloitte.
Insights

Sign up for Deloitte Insights updates at www.deloitte.com/insights.

Follow @DeloitteInsight

Deloitte Insights contributors

Editorial: Abrar Khan, Arpan Kr. Saha, Aparna Prusty, and Emma Downey

Creative: Sonya Vasilieff and Ayushi Mishra

Audience development: Kelly Cherry, Maria Martin Cirujano, and Nikita Garia

Cover artwork: Sonya Vasilieff