Google

Secure, Empower, Advance

How AI Can Reverse the Defender’s Dilemma

February 2024

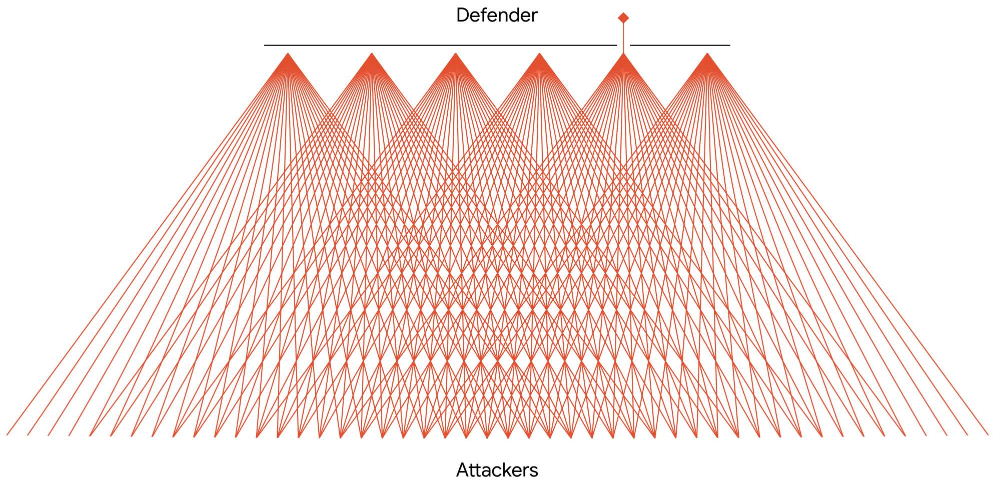
Table of contents

Section 1

Executive Summary

The invention of the internet unlocked unprecedented innovation and economic opportunity

But while the internet's core technologies fostered rapid innovation, interoperability, and the free flow of information, these technologies were not designed with security in mind. The explosive growth of the digital domain on top of this foundation created an environment conducive to a wide range of malicious behaviors. Attackers possess inherent advantages in cyberspace: they can choose from a wide variety of targets and need only succeed once, while defenders must protect an increasingly complex terrain and need to be successful at all times. This dynamic, referred to as the "**Defender's Dilemma**," has plagued organizations and users for decades.

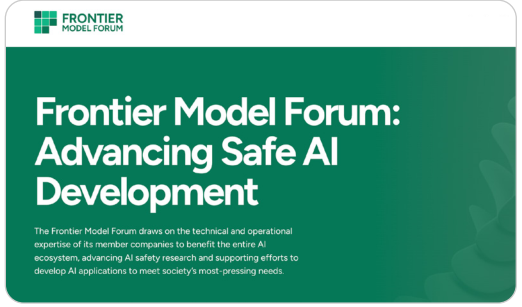


We believe AI affords the best opportunity to upend the **Defender's Dilemma**, and tilt the scales of cyberspace to give defenders a decisive advantage over attackers.

Enter artificial intelligence (AI)

The advent of AI is already reshaping the digital world. We believe AI affords the best opportunity to upend the Defender's Dilemma, and tilt the scales of cyberspace to give defenders a decisive advantage over attackers. AI will enable us to effectively cope with the complexity of our digital world and can help turn every organization into a competent defender. This will create new paradigms for security and software development, and correct many of the asymmetries in capabilities and resources that give attackers an edge online.

Digital enterprises have learned hard lessons about how to secure computers and systems, attempting to compensate for the fundamental flaws in the internet. Now, we have the chance to design AI security tools the way we want them to be, built securely from the start. To date, there has been a strong and appropriate focus on addressing potential future risks from AI. We have seen governments take important steps together with companies and other civil society stakeholders to address and mitigate these risks. It's why Google co-founded the Frontier Model Forum (FMF) to advance AI safety research, and created the Secure AI Framework (SAIF) as a vehicle to collaborate on best practices for securing AI systems.


FRONTIER MODEL FORUM

Frontier Model Forum: Advancing Safe AI Development

The Frontier Model Forum draws on the technical and operational expertise of its member companies to benefit the entire AI ecosystem, advancing AI safety research and supporting efforts to develop AI applications to meet society's most-pressing needs.

Figure 1

Frontier Model Forum (FMF)

But as we described in The AI Opportunity Agenda, to fully harness AI's transformative potential, we need a broader discussion about steps that governments, companies, and civil society can take to realize AI's promise. We must focus not only on the harms we want to avoid and the risks we want to mitigate, but on the potential outcomes we want to achieve. This is true across the board for AI, but particularly for AI's ability to revolutionize security.

To achieve this goal, this paper makes **three key recommendations**:

- **Secure AI from the ground up.** AI-powered security must sit on a trustworthy foundation for the technology to correct some of the original shortcomings of our digital domain. Applying the lessons learned from decades of cybersecurity is vital during the excitement of this moment. Recent policy and industry efforts have focused on mitigating foundation model risks, but models are only one part of the systems that users and enterprises will interact with. Secure-by-design principles need to infuse the lifecycle of the technology at all layers of the stack. And to ensure the technology can be trusted to deploy at scale, we must collaborate on developing new guardrails for autonomous cybersecurity.

- **Empower defenders over attackers.** Our societies need a balanced regulatory approach to AI usage and adoption to avoid a future where attackers can innovate but defenders cannot. AI governance choices made today can shift the terrain in cyberspace in unintended ways. There are a number of actions we can take today to ensure we maximize the technology's utility for defenders, while minimizing malicious use. While AI risk management is critical, certain policy approaches — such as those which limit AI usage in critical infrastructure, or allow users to opt-out of AI security functions — will bind the hands of cyber defenders but leave attacker use of the technology unconstrained. We can work together to give defenders the upper hand — such as by pooling security-relevant datasets to ensure defenders have access to better models than attackers.

- **Advance research cooperation to generate scientific breakthroughs.** The research community must play a central role in enabling new paradigms for security and software development. This includes testing and evaluating new security technologies, assessing and prioritizing risks, and introducing new innovations to help eliminate entire classes of threats. While existing publications tend to focus on demonstrating attacks on or using AI, we should prioritize research into building defenses against or with AI.

Capturing the opportunity to shape the direction of AI-powered security will take bold investments and cooperation across governments, industries, and civil society. Reversing the Defender's Dilemma is an ambitious goal, and achieving it is by no means assured. Attackers will work just as hard to undermine these efforts. But this is why bold and timely action is needed today.

Section 2

Evolving Threat Landscape

Despite years of intense policy focus and industry investment, attackers pose greater risk to societies than ever.

As an industry-leading external threat intelligence service provider, we see threats continue to grow on a global scale. Google Threat Intelligence teams now track hundreds of threat actors and thousands of malware families, and have generated tens of thousands of threat intelligence reports.

Through this ongoing, frontline engagement, our experts anticipate **three key trends:**

1. Both criminal and state sponsored threat actors are continuing to professionalize operations and programs.
2. Offensive cyber capability is now a top geopolitical priority for most governments.
3. Threat actor groups' tactics now regularly evade "standard" controls.

We've also observed unprecedented developments like the Russian invasion of Ukraine marking the first time cyber operations played a prominent role in war.

The threat landscape remains dynamic and complex, and we expect these trends to continue throughout 2024 and beyond.

To help organizations better prepare, we recently outlined key cybersecurity forecasts for 2024 on topics like adversary use of AI and zero-day vulnerabilities. We also expect financially motivated threat actor groups to continue to deploy ransomware and extortion tactics to take advantage of victim organizations and drive up costs for everyone. By 2028, cybercrime will cost an estimated 13.8 trillion dollars worldwide.

We assess with high confidence that China, Russia, North Korea, and Iran (the "Big Four") will continue to pose significant risks for defenders across geographies and sectors. China in particular has been investing heavily in using AI for offense and defense, and engaging in intellectual property and personal data theft to enhance AI competition with the United States.

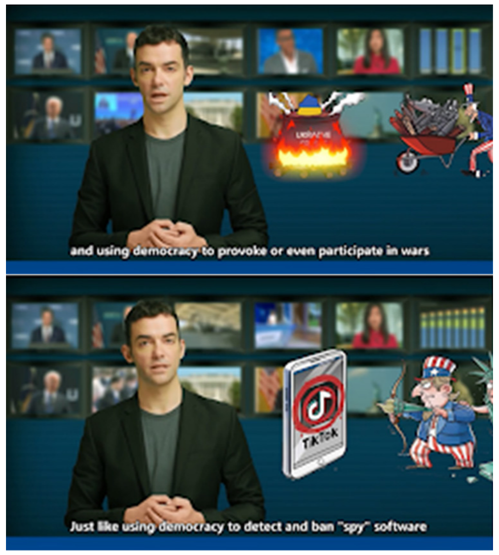

and using democracy to provoke or even participate in wars
Just like using democracy to detect and ban "spy" software

Figure 2

Screenshots from video containing AI-generated "news presenter" promoted by DRAGONBRIDGE, likely created using a platform offered by D-ID

These trends are equally important in the context of AI. Since at least 2019, we've tracked threat actor interest in, and use of, AI capabilities to facilitate a variety of malicious activity. Based on our own observations and open source accounts, adoption of AI in intrusion operations remains limited and primarily related to social engineering. In contrast, information operations actors of diverse motivations and capabilities have increasingly leveraged AI-generated content — particularly imagery and video — in their campaigns, at least in part because of the readily apparent application of AI to disinformation. Ongoing investments in forensics and in technologies like detection, watermarking, fingerprinting, and signed metadata will over time raise barriers against these malicious uses of AI.

Social Engineering

Generative AI and large language models (LLMs) will be utilized in phishing, SMS, and other social engineering operations to make the content and material (including voice and video) appear more legitimate. For example, in March 2023, multiple media outlets reported how a Canadian couple were scammed out of $21,000 when someone using an AI-generated voice impersonated their son as well as their son's representing attorney for allegedly killing a diplomat in a car accident.

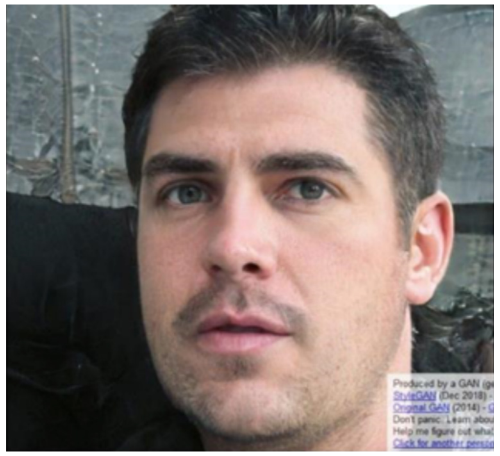


Figure 3
An AI-generated image used as a profile photo by a persona in a pro-Cuban government network displayed a text box showing that the image was generated using the website thispersondoesnotexist.com

Information Operations

Attackers will use clever generative AI (gen AI) prompts to create fake news, generate fake phone calls that actively interact with recipients, and produce deepfake photos and videos based on gen AI-created fake content.

For example, in March 2023, DRAGONBRIDGE unsuccessfully leveraged several AI-generated images in order to support narratives negatively portraying US leaders. One such image used by DRAGONBRIDGE was originally produced by the journalist Eliot Higgins, who stated in a tweet that he used Midjourney to generate the images, suggesting that he did so to demonstrate the tool's potential uses. We judge that DRAGONBRIDGE has not gained traction with their campaigns.

As AI technology evolves, we believe it has the potential to significantly augment malicious operations in the future, enabling threat actors with limited resources and capabilities, similar to the advantages provided by exploit frameworks including Metasploit or Cobalt Strike. As a result, we expect to see more adversary use of AI tools over time. Government and industry must scale to meet these threats with strong threat intelligence programs and robust collaboration. These efforts will help stay abreast of the threat but will not fundamentally change the current dynamics which plague defenders.

Section 3

The Defender's Dilemma

Responsible actors — including IT professionals, developers, cyber defenders, and even everyday users — face a seemingly impossible task

Defenders are outmatched by attackers. This "Defender's Dilemma" has many well-known (but poorly constructed and often over-simplified) phrases associated with it:

> *"There are two types of companies: Those who know they've been hacked, and those who don't."*

> *"On a long enough timeline, an advanced threat actor will always achieve their objectives."*

> *"Defenders have to be right every time. Attackers only need to be right once."*

How did we get here?

The answer lies in the core attributes of the internet and the economic systems that have sprung up around it. The internet's core protocols enabled rapid innovation, interoperability, and the free flow of information. Yet, these same components created an environment that is vastly distributed and complex. As economic functions and nationally-important datasets began to be connected, threat actors connected as well, engaging in a range of malicious behaviors, from espionage to cyber crime. This transition removed geographic barriers of entry to reach potential victims, and defenders have been playing catch-up ever since.

Three factors in particular contribute to the structural conditions underlying the "Defender's Dilemma":

The internet was designed to move information, not protect it

Security was not a core requirement of the original internet technology stack. The internet's founding protocols — including TCP/IP, DNS, and BGP — were optimized for resilience and reliability, ultimately ensuring that data could be routed where it needed to go even when under nuclear attack. The core stack did not emphasize concepts like identity, authentication, and authorization, which are important precursors for security. Some of these foundational protocols have been revamped to add security features, but these processes occur on generational timescales, and are fraught with the potential to break compatibility with older versions. If we redesigned online infrastructure from the ground up today, we would likely make very different design choices and tradeoffs.

Our digital ecosystem grows more complex each year

Perhaps the greatest feature of the internet is its interoperable architecture. Users can connect different kinds of technologies together over the web's foundational protocols, and have them work together. Meanwhile, as software has become more complex, developers have created abstractions to manage it. In other words, humans add more — yet ostensibly simpler — layers that hide the complexity, versus eliminate it. But the complexity isn't gone, it is just hidden. Being designed by humans, each of these layers of abstraction comes with vulnerabilities that stem from the interplay between them. This can come in the form of developers unintentionally introducing vulnerabilities into the codebase, network administrators misconfiguring the network, or users clicking the wrong link. The problem here is not with the human user — it's with the system itself.

Our digital world is now a vast mosaic of software and services that grows more complex each day. Complexity is not inherently bad: complex systems knit together different datasets and services to provide value to people. But complexity is hard to manage, and unmanaged complexity introduces systemic risks. Each year, security breaches occur because attackers induced software to perform in ways the developer did not expect, or compromised devices the network operator did not realize were connected to the network. Absent intervention, the rise of AI will contribute to this problem, as the technology helps developers create and manage more software.

Structural asymmetries advantage attackers, and hinder small defenders

Defenders are spread thin throughout the ecosystem, while attackers can focus their efforts and benefit from a number of structural advantages.

Resource Asymmetries

Millions of organizations around the world hold some value to attackers, such as money, data, or something else, like procuring infrastructure to launch more attacks. Most of these organizations do not have any IT personnel, let alone cybersecurity experts. What hope does a small business like an accounting or construction firm have of defending against a determined attacker?

Attention Asymmetries

Even well-resourced companies face challenges, because as the company grows, so does its attack surface. Given the complexity of modern systems, defenders can quickly become overwhelmed with triaging security alerts. The breadth of attack techniques and threat actors means defenders must prepare for dozens of scenarios. Defender burn-out is a chronic problem. By contrast, attackers have the luxury of massing their efforts against a single organization. They can choose targets of opportunity (e.g. the least well-defended organizations), or patiently wait for defenders to make mistakes. Defenders, on the other hand, must always stay vigilant.

Information Asymmetries

Attackers can find vulnerabilities in common software, which they can keep secret and exploit at will. These zero-day vulnerabilities can be used for years without the defender community discovering them. Further, over 40% of the zero-days we discovered in 2022 were variants of previously reported vulnerabilities.

Our industry has made significant incremental investments and progress to make exploitation more difficult and cybersecurity easier to adopt. Yet, no new security innovation or policy initiative has fundamentally reversed the Defender's Dilemma. Classic approaches to cybersecurity from legacy vendors are reactive and still too manual and error prone. Organizations must stay abreast of new techniques and vulnerabilities, and constantly prioritize mitigations as soon as they are released. This is becoming unsustainable as the threat surface continues to grow and attackers increasingly look to AI to scale their operations.

Moving our ecosystem forward onto a more sustainable, secure path will require a new approach, one which addresses the growing complexity of our online world and reverses the asymmetric advantages that attackers hold over defenders.

Section 4

Enter Artificial Intelligence

Over time, AI will be the biggest technological shift we see in our lifetimes.

It’s bigger than the shift from desktop computing to mobile, and it may be bigger than the internet itself. It’s a fundamental rewiring of technology and an incredible accelerant of human ingenuity.

Sundar Pichai
CEO of Google and Alphabet

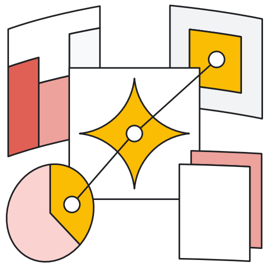
Over the last year, the rise of generative AI has wowed us with tools that can write poems, create lifelike images from text prompts, and provide thoughtful responses to questions both profound and practical. But AI's utility reaches far beyond chatbots.

AI is optimizing complex logistics for global businesses and unlocking the human genome to drive medical breakthroughs. As with many other domains of human endeavor, AI will completely transform online safety and security.

To aid cyber defenders, AI is already analyzing vast amounts of data to identify anomalies; automating routine security functions; and serving as a helpful assistant to human analysts triaging and actioning alerts. Over time, AI will help us go much further. As the field advances, autonomous agents will begin to knit other AI systems together. Advances in AI can lead to self-healing software and networks by learning from trends in attacker behavior, using them to identify vulnerabilities, generating safe code and configuration fixes, and deploying them to production rapidly.

These advancements are possible due to **AI's unique attributes,** which will disrupt the field of cybersecurity along with many others.

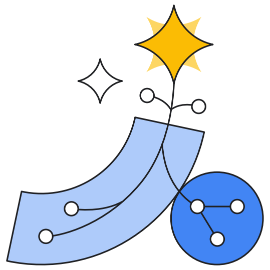
Reasoning



Much of the power of AI stems from its ability to rapidly analyze information, draw logical conclusions, and make decisions. AI can perform a variety of reasoning tasks, including deductive reasoning (generating inferences based on pre-determined rules), inductive reasoning (making probabilistic inferences based on observations), and abductive reasoning (forming inferences from known facts). Reasoning enables AI systems to make decisions, and thus perform tasks. As AI grows more capable, the tasks it can perform will become more useful and complex. This will be transformative for cybersecurity, as defenders struggle both with the amount of tasks that must be performed in a modern environment, and the difficulty of certain tasks. AI can already perform high confidence analysis on complex datasets far faster than humans. Consider malware detection — Google Cloud's VirusTotal has been applying AI to reason about unknown files and determine whether they are malicious. Here, AI is able to crawl through millions of lines of code to a painstaking level of detail, and reason about whether some part of that code is used to break into systems, using algorithms trained on VirusTotal's vast historical dataset of malware. In November, VirusTotal reported that AI excels in identifying malicious scripts, particularly obfuscated ones, achieving up to 70% better detection rates compared to traditional methods alone. AI demonstrates enhanced detection and identification of scripts exploiting vulnerabilities, with an improvement on exploit identification of up to 300% over traditional tools.

Learning

AI's ability to reason effectively stems from its ability to learn. Machine learning enables an AI system to improve its performance on a given task without being explicitly programmed for every specific scenario. Learning is performed through a variety of techniques. Some models will learn by analyzing large, unstructured datasets, while others are trained on smaller, high-quality datasets. Learning is vital for systems to be valuable in complex, dynamic fields — like cybersecurity. Attackers are constantly evolving their techniques and tools, and defenders must constantly adapt to keep up. AI systems are a game changer for identifying new threats while minimizing false positives. For example, Google Cloud's Anti-Money Laundering AI (AML AI) product uses machine learning across millions of financial institutions' banking transactions to identify suspicious financial activities. The tool is already being used to detect 2–4 times more true positives while reducing alert volume by 60%.

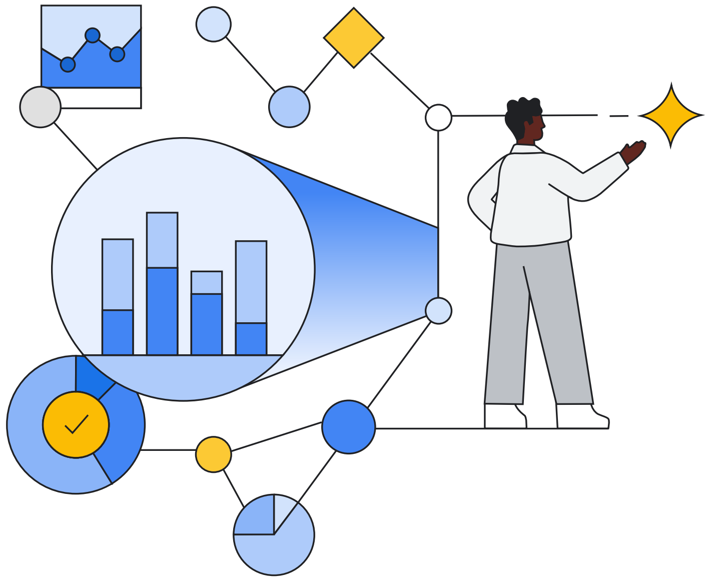


Speed

"Defenses must move at the speed of cyber" is yet another trope associated with the Defender's Dilemma, but this has never been possible until AI. Humans are inefficient at performing many routine cybersecurity tasks, but AI systems operate at machine speed. The ability to quickly evolve defenses, apply patches, and detect attacks faster makes all the difference in reducing attacker "dwell time" and keeping them out in the first place. AI can both make decisions on its own in real-time, as well as help humans make important decisions faster. Generative AI is already proving valuable in helping security analysts investigate new issues faster. Google Cloud's Gemini, for example, can help analysts by rapidly searching vast datasets based on natural language requests; automatically summarizing case data and alerts; and improving response time by recommending next steps for remediation. Internally, our Detection & Response teams have seen a 51% time savings and higher quality results in incident analyst output using generative AI.

Scale

AI is unique in its ability to handle diverse data at scale, quickly and autonomously analyzing, sorting, and making sense of data sets far larger than any human could handle. For example, our AI-powered Enhanced Safe Browsing capability examines billions of URLs entered into Chrome against millions of known malicious web resources, and sends more than three million warnings per day to users. AI systems can identify patterns and correlations, detect anomalies, and create predictions from these immense and ever-growing datasets. AI can handle both specific, structured data like telemetry and unstructured data like images or videos, potentially integrating these capabilities into a holistic approach to security at scale. This means that as more data becomes available, AI can continue to learn and improve to handle the current conditions rather than reacting to yesterday's attack. The scale of our digital security problem — every corner of the internet, and every corner of an organization's digital infrastructure — means that security is no longer achievable at human scale. But with assistance from AI, organizations can automate simple and complex tasks, and do them at almost any required scale.



AI Cybersecurity Use Cases

AI is not one particular field or discipline, it is many.

- AI can be used to predict the outcome of a given input, often assigning a confidence score.
- Other AI systems can generate content for consumers or enterprises.
- Large-language models get much of the attention today, and are trained on very large datasets to enable general purpose natural language processing and generation.
- Other models, such as expert models trained on domain-specific datasets, can be just as useful for solving narrower problems.
- Some AI systems can take images as inputs, while others use text or code. Multimodal systems can analyze multiple forms of inputs.
- In time, we will increasingly see AI act as supervised agents, able to perform a multitude of tasks on our behalf with less need for direct human interaction.

Different AI capabilities can help defenders with different cybersecurity tasks:

Summarize

Complex data, intuitively accessible: Provide quick and simple ways to search through intelligence and provide explanations so users understand attack exposure, impacted assets, and mitigations.

Capabilities

- Concisely explain behavior of suspicious scripts
- Summarize relevant and actionable threat intelligence and reports
- Summarize case investigations
- Summarize vulnerability reports

Classify

Critical insights, readily surfaced: Reason about *criticality and risk*, so valuable information can be shared more quickly; Quickly understand implications of large-scale events to drive investigations forward; Automatically update cloud security policies to keep pace with known threat information.

Capabilities

- Classify malware
- Identify security vulnerabilities in code
- Categorize and prioritize threats
- Detect unusual and malicious events
- Run attack path simulations
- Monitor the performance of controls and assess early risk of failures

Create

Specialized syntax, instantly translated: Simply provides the parameters to create a query, detection, or rule — without the need to be an expert in specialized security languages such as YARA. Generate complex queries and transform the task of writing an effective detection from minutes or hours of work to seconds.

Capabilities

- Generate queries from natural language
- Create detection rules
- Generate security orchestration, automation and response playbooks
- Generate identity and access management rules and policies

Section 5

Roadmap to Digital Security

AI's core attributes will disrupt cybersecurity

Our task is to ensure this disruption can maximize benefit to users and organizations while minimizing harms. AI, like most other useful technologies, can be used for malicious purposes. A system that can find vulnerabilities for defenders to fix can also find vulnerabilities for attackers to exploit. Without careful intervention and close cooperation to direct the technology's evolution and use, attacker use of AI could result in yet another arms race which merely projects the Defender's Dilemma deep into the future.

But if we seize the moment, we believe AI can usher in **two fundamental paradigm shifts** which address the root causes of cyber insecurity:

1. Using AI to understand and help us manage the complexity that generates so much vulnerability in the digital domain.
2. Using AI to uplevel all users of digital technology to be a competent defender, and in select cases, eventually move from assistive to autonomous.

The internet connects tens of thousands of small organizations with little or no knowledge of cybersecurity. AI can put a capable security expert in each of them.

In short, AI may make attackers better, but the gains will not be nearly as great as those felt by democratizing security expertise for everyone.

Abstracting Away Complexity, Securely

AI can handle complexity in a way humans cannot. Whether it is reasoning about a complex codebase, or the complex interactions between systems at global scale, we think AI can be evolved to address the complexity crisis at the heart of the Defender's Dilemma.

AI systems can perform tedious tasks at scale, which allows them to evaluate, generate insights, and make decisions about very large datasets and action spaces. First and foremost, this will help humans understand, and then optimize, the software they are building and the networks they are tasked with defending.

In short, AI can help us understand the way technology truly works, rather than the way we think it works.

For example, generative AI is being used today to quickly summarize the functionality of files in plain language — providing value both to engineers seeking to incorporate new functionality in their environment, and to help security analysts uncover malicious functionality within an otherwise harmless program. Countless security incidents result from mistaken assumptions and errors introduced by humans while managing complex systems — AI can help us correct them.

Embedding AI into software development — to better understand the software we are building, and align software with secure principles — is perhaps the most significant step here. Security takes the biggest leap forward when we design products that eliminate entire classes of vulnerabilities. This starts with designing build processes which are secure-by-default to manage complexity for developers and limit the ability for humans to insert errors into the codebase or system. AI has the potential to guide developers to make more secure choices; review architectures to more consistently enforce security principles; and monitor development environments for compliance.

Early tests reveal AI's potential to fix vulnerabilities in code. We harnessed our Gemini model to successfully fix 15% of bugs discovered by our sanitizer tools during testing, resulting in hundreds of bugs patched. Given the large number of sanitizer bugs found each year, this seemingly modest success rate will save significant engineering effort. We expect this success rate to continually improve and anticipate that LLMs can be used to fix bugs in various languages across the software development lifecycle. In the future, AI systems may help us rewrite legacy components in memory safe languages, speeding along an evolution to eliminate the class of bugs responsible for the most severe vulnerabilities in the digital domain.

Deployers and users of technology will benefit just as much as developers, and the benefits here are being realized today on a far greater scale. AI is already being used to identify misconfigurations, tailor permissions to users based on their demonstrated access patterns, optimize traffic flows, and more. We can also identify anomalous activity based on a previously observed baseline.

Just as AI can help us understand our own technology, it will help us understand adversaries. Foundational large-language models get most of the attention, but narrower "expert systems" will be just as impactful for this. Expert systems are built on carefully curated bodies of knowledge and are designed to address problems within specialized domains. One example is malicious email detection. In November, we announced RETVec, a new multilingual neuro-based text processing model. Compared to large-language models, which can often contain tens of billions of parameters, RETVec is micro-sized at only 230k parameters. Yet, deployment in Gmail improved spam detection rates over the baseline by 38% and reduced false positives by more than 19%.

Vulnerability Discovery Deep Dive

Flipping Information Asymmetries to Aid Defenders

Today, a single person acting alone can develop an exploit that poses extreme systemic risk — and keep that information private.

Lone actors can (and do) find dangerous vulnerabilities before even the most sophisticated technology companies. These so-called zero-day vulnerabilities act like a skeleton key to enable access to systems undetected by defenders. As long as the knowledge of the vulnerability remains secret, defenders have little chance to stop these attacks.

We are seeing early signs that AI technologies will be able to discover exploitable vulnerabilities in code far more comprehensively than humans. As public interest technologist Bruce Schneier explains, "[g]oing through software code line by line is exactly the sort of tedious problem at which machine learning systems excel, if they can only be taught how to recognize a vulnerability."

Some commenters are concerned that breakthroughs in this area will exacerbate zero-day exploitation in the wild, but **we think the opposite is true**: advances in AI-powered vulnerability and exploit discovery will benefit defenders more than attackers.

What if AI can begin to surface bugs that only attackers and extremely well-resourced security teams can find today?

What if AI begins to outstrip human capabilities altogether?

Empowering defenders with this capability has the potential to revolutionize the field. Throughout our history, discovery and public disclosure of vulnerabilities consistently produces digital products and services that are more secure, reliable, and trustworthy. This is the mission of our Project Zero team — to make zero-day hard by driving awareness and ecosystem-wide mitigations. AI-powered vulnerability discovery could help in a similar way, on a vast scale. Embedding this technology within build systems can drive not just exploit mitigation but prevention of entire classes of bugs. The impact will be to reduce or even flip the information asymmetries that currently favor attackers — giving defenders the commanding view of system weaknesses.

Government and industry are beginning to rally around this challenge. DARPA's AI Cyber Challenge will explore using AI technologies to automatically find and fix vulnerable open source code. Imagine a world where the open-source projects upon which our digital world runs can all benefit from the level of maintenance and hardening they deserve, through the power of AI. This may not be as far-off as it seems.

Scaling Security Expertise

The modern technology environment places far too much responsibility on organizations and people who are not security specialists. As the US National Cyber Strategy puts it, "Today, end users bear too great a burden for mitigating cyber risks[...]Our collective cyber resilience cannot rely on the constant vigilance of our smallest organizations and individual citizens."

AI has the potential to relieve the burden on these end users and make organizations more capable in cyber defense.

Early studies show that inexperienced workers stand to gain the most from AI, while benefits to skilled workers will be more incremental — an effect some are calling "the great equalizer." This effect will be felt in cybersecurity, as organizations without any cybersecurity expertise will be able to leverage AI to enable a baseline security posture. This can radically reshape the balance of power online, with far less "low-hanging fruit" for attackers to prey upon.

In the future, organizations will be able to benefit from AI security experts, which will both empower humans to be more effective while performing challenging tasks, and potentially relieve them of entire classes of toilsome cybersecurity functions altogether. This in many ways accelerates a trend that started with cloud adoption, where security tasks (e.g., patching the infrastructure) are shifted from end users and organizations onto the cloud provider's security specialist teams. AI agents will be able to perform as cybersecurity experts within an organization by linking together more narrow AI uses within a general autonomous cyber defense framework. These systems can help with a wide variety of tasks, like threat management, continuous monitoring and incident response.

Gen AI has already unlocked the ability of humans to interact with systems through natural language and accomplish tasks that seemed impossible previously. Tools built upon security-specific large-language models, such as Google Cloud's SecLM, can help analysts search billions of security events and interact conversationally with the results, ask follow-up questions, and quickly generate detections — all without learning a new syntax. They can translate complex attack graphs to human-readable explanations of attack exposure, generate summaries of impacted assets and provide recommended mitigations. Embedding these models in front-line tools like Gemini can help even non-experts detect, investigate, and respond to cyberthreats with confidence.

Attackers will be upleveled by the technology, but the aggregate effect will aid the defense far more. Each malicious actor, by its nature, possesses at least some capability. The same cannot be said of each organization's capability to defend. The internet connects tens of thousands of small organizations with little or no knowledge of cybersecurity. AI can put a capable security expert in each of them. In short, AI may make attackers better, but the gains will not be nearly as great as those felt by democratizing security expertise for everyone.

How Mandiant Consultants and Analysts are Leveraging AI Today

Today, Mandiant is leveraging generative AI in bottom-up use cases to help identify threats faster, eliminate toil, and better scale talent and expertise that increase the speed and skill we bring to serving customers.

Last year, we highlighted several examples across Mandiant's consulting and analysis teams that used Gemini within their workflow. **NOTE:** No client data is entered into Gemini and the output is reviewed before any implementation.

EXAMPLE: Analyzing an Adversary's Smart Contracts

Smart contracts are effectively computer programs stored on blockchains such as Ethereum. While there are many legitimate roles for smart contracts, threat actors have utilized them to serve as the foundation of malicious projects, for theft of cryptocurrency assets, and to obfuscate movement of funds. Mandiant analysts tracking criminal use of cryptocurrencies are faced with a daunting task: tracking and understanding thousands of smart contracts with varying levels of complexity and functionality. Unlike most legitimate smart contracts, threat actors do not publish the source code of their projects, leaving analysts with only a contract's non-human-readable form of computer code called bytecode.

Analysis of a threat actor's smart contract involves analyzing oft-obscurely named functions in the bytecode. This can quickly become a tiresome, complicated task — particularly if an analyst is not well versed in Solidity, a programming language used to develop smart contracts. In recent analytic efforts, Mandiant analysts used Gemini to assist with analyzing threat actors' smart contracts. Not only can Gemini describe a function's purpose, it can also provide an easy to understand line-by-line commentary of the bytecode, helping analysts prioritize their analysis according to each function's role.

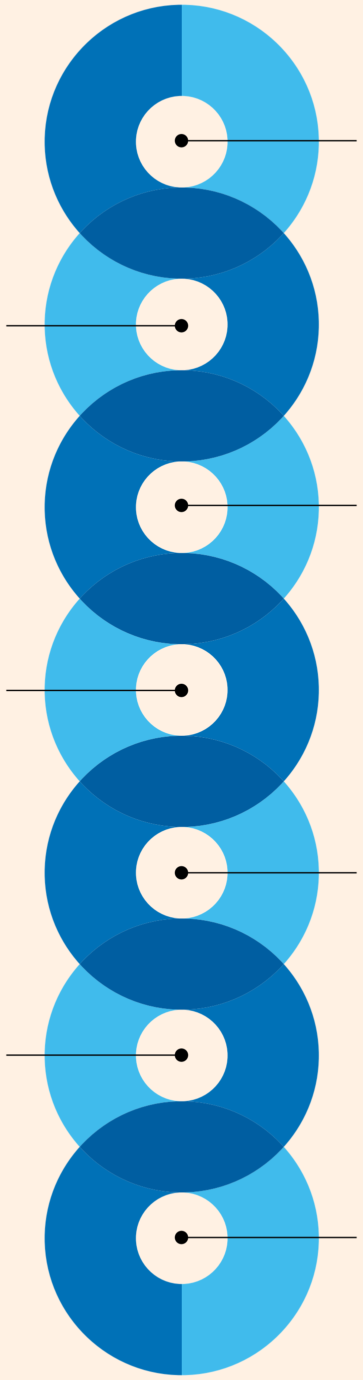
This capability has been particularly useful investigating adversaries such as the North Korea-affiliated group we call UNC4469, which has used thousands of smart contracts to steal funds.

Moving towards an Intelligent Digital Immune System

As we've discussed above, specific advancements will lead to revolutions in how software is securely developed, deployed, run and managed. Over time, AI has the ability to merge and automate the feedback loop between these functions, creating an AI-based digital immune system to protect organizations. Expert domain models will begin to help us manage complexity in our computing environment, and agents will help people of various skill levels make use of them. Each new exploitation attempt will generate real-time learning that can be shared in real time across the cloud. This will drive rapid adaptation not just in threat detection, but in coding, deployment and runtime management practices as well.

> Over time, AI has the ability to merge and automate the feedback loop between these functions, creating an AI-based digital immune system to protect organizations.

An Intelligent Digital Immune System



Threats

AI-driven threat intelligence systems monitor attackers trends

Vulnerability Discovery

Vulnerability discovery systems learn from attacker trends, driving discovery of new vulnerabilities and misconfigurations

Code and Configuration Safety

CI/CD systems learn from new vulnerabilities and misconfigurations and update secure coding and deployment guardrails for developers and administrators

Secure Code Generation

AI can propose new code and secure-by-default configurations to patch weaknesses

Automated Updates

AI can test and deploy new patches and configuration changes

Detection and Incident Response

Systems can learn from baseline endpoint telemetry and user behavior to detect threats in the environment, summarize alerts and incidents for analysts, and propose remediation steps

Continuous Monitoring

AI can continuously monitor system performance and controls posture, and make recommendations

AI experts can knit together these systems and workflows to create self-healing networks

Current AI systems can perform some of these tasks, but complete self-healing networks are not a reality. The path will be long, and we can't predict all the ways this technology will evolve, however it is helpful to set a marker and consider how the technology must advance to meet the goal:

1. AI must be built with strong secure-by-design fundamentals and deployed and run in a secure-by-default manner.
2. AI must manage immense complexity while promoting high-quality, reliable answers. Ultimately, we should move towards formal methods where we have more confidence in AI enabled defenses.
3. AI must knit capabilities together and generalize them to provide general security expertise that is transferable to new and unseen domains.
4. To ensure broadest impact, AI must have methods to programmatically interface with existing systems, protocols, and data. This does not mean every organization needs to acquire a suite of standalone AI solutions. AI should be baked into modern platforms (e.g., devices, browsers, cloud platforms) to achieve the greatest benefit for the most users.

Section 6

Capturing the Opportunity

This is a pivotal moment.

Now is the time for us to come together to tip the scales in favor of defenders.

AI represents the greatest opportunity since the internet's creation to reverse the Defender's Dilemma, but it is not a silver bullet. Its effectiveness for cybersecurity depends on factors such as the quality of the AI systems, the data they are trained on, and the extent of deployment. As AI evolves, the asymmetries in cyber defense and offense will continue to shift, and it will be an ongoing challenge for organizations and governments to adapt their cybersecurity strategies accordingly. To ensure that these asymmetries tip in favor of defenders, we need a bold research and policy agenda to unlock the science and create structural conditions to provide maximum leverage for defenders and limit the potential for malicious use. This agenda includes **three broad pillars:**

Secure AI from the ground up

AI's potential to reshape the internet also offers the chance to fix some of its original flaws.

Security was bolted-on to the internet after the fact — we can do better this time. But as we've learned from prior waves of technology, the benefits of new innovations will not come automatically. People must trust the technology before it will be adopted at scale and used in high-impact settings. This is doubly important for security technologies, which, by their nature, perform important functions and require sensitive access to operate, thus making them an inviting target in their own right for attackers. AI security technologies must be secure-by-design, and deployed in a secure-by-default manner, or they will become vectors for vulnerabilities like any other technology, fueling the trend we are seeking to stop.

AI security technologies must be secure-by-design,and deployed in a secure-by-default manner, or they will become vectors for vulnerabilities like any other technology, fueling the trend we are seeking to stop.

Prioritize holistic security and resilience of AI systems

Much of the policy attention to date has been on the long-tail safety risks presented by AI models, which has naturally led to an emphasis on securing the model weights and ensuring model outputs are rendered safe. A number of new initiatives are driving this area forward, including the creation of AI Safety Institutes in the US and UK, as well as industry groups like the Frontier Model Forum.

Yet, users do not interact with models directly. Models are embedded within technology products, consisting of hardware and software, some of it cutting-edge, but often legacy as well. Attackers choose the path of least resistance to accomplish their end goals, and we believe this will seldom require attacks on the model itself. We cannot ignore the more "mundane" security risks which can be introduced throughout the lifecycle of AI systems (from pre-training through to deployment and run time) and at all layers of the stack (hardware, operating systems, protocols, APIs, etc). Users today are already suffering from lack of investment in more traditional areas of security, such as ensuring systems are patched. This is why we launched SAIF to collaboratively build best practices for securing AI systems. AI developers should apply leading security best practices, including using hardened infrastructure for training, employing software supply chain security best practices, and ensuring the developer's corporate environment is secured from insider risks and account compromises. These secure AI principles should be embedded in new policy initiatives to update procurement guidelines and critical infrastructure regulations for the AI era.

Build a risk-based approach for autonomous cyber defenses

The scale of online threats has already outstripped human capacity. Leveraging autonomous capabilities for defense will be mandatory in the future. Many security tasks have been safely handled by AI systems for years, but as these systems grow more powerful we will need guardrails on their usage to ensure they are aligned with our values. AI systems must be auditable so people can ensure lawful, appropriate, and proportionate actions are taken. We will need to ensure effective human oversight, allowing human operators to redirect or stop the system if needed. That said, these requirements for humans-in-the-loop should be risk-based to ensure AI systems can provide maximum value to defenders. These are the kinds of challenges FMF was created to address, and we will lead efforts to develop these guidelines in coordination with partners.

Promote skilling opportunities for AI and cyber

Ensuring that AI is built securely, and used to advance security for all, starts with a capable workforce. Government and industry must build on efforts to expand pathways into careers in both AI and security. At Google, we are creating new pathways enabling cybersecurity and AI careers for all, through investments in cybersecurity clinics around the world, certificates for early entrants, support for research in the field, and creating community partnerships.

Collaborate on best practices for AI-powered security

Many new initiatives have sprung up in the past year to manage the risks of AI. We think it is time for new partnerships to focus on how to use AI to manage broader security risks in the digital domain. While initiatives such as the US AI Executive Order contain new steps to explore AI's use for security, these must be built upon. To help support this effort, we're partnering with industry, academia, and others to advance best practices and introduce new tools across all six SAIF elements. We encourage every organization to work together to implement SAIF and build on this momentum to secure AI systems.

Empower defenders over attackers

Attackers are already innovating with AI. While major efforts to drive AI responsibility and accountability are needed, we can't lose sight of the incredible opportunity AI presents for cybersecurity. It is vital that approaches to AI governance do not tie the hands of defenders at a time when attacker experimentation is accelerating.

Ensure the Best Models are Built for Defenders

The effectiveness of AI is based in large part on their underlying models, and model effectiveness is based in large part on the quantity and quality of the data used to train them. Thankfully, we think defenders have a (tenuous) advantage today. Models built for defenders by defenders are already benefiting from a vast quantity of security-relevant data held by cybersecurity and platform companies, enterprise organizations, and governments. While more sophisticated threat actors potentially have access to their own private datasets that can rival a given defensive model developer, their numbers are few, and none can rival the combined efforts of the cybersecurity community. A single cybersecurity company can build a model that can learn in near real-time from hundreds (or thousands) of customer environments around the world. However, this advantage is fragile. Some attackers will have better models than a given defenders, and attackers can subvert or steal models. Our task is to ensure that the defensive community has an information advantage which can be converted into a model advantage; that this advantage is scaled as far across the ecosystem as possible; and that our foundational approach to AI safety and security ensure these models cannot be misused.

- **Preserve the ability to train models on publicly available data**
  The most powerful and effective models today are trained on large, publicly available datasets, and then enriched with private datasets and various fine-tuning techniques. Some policy proposals have contemplated prohibiting the use of large public datasets. This would create negative unintended impacts on cybersecurity. Barring companies from using a valuable resource that is publicly available will create a reality where the only actors using them are those who are not bound to the rule of law.

- **Share and collaborate on security training datasets**
  The cybersecurity community has launched (and re-launched) various information sharing initiatives over the past two decades. Sharing security data to inform AI model development is one clear area where information sharing would have significant value, driving the creation of better models for defense. Governments should consider ways to foster partnerships for the creation of better security domain models — including by publicly releasing their own useful datasets.

Do not require opt-outs for AI security functions

Some policy proposals would allow users to opt-out of AI-powered decision-making tools. While these proposals may be intended to protect individual rights, policymakers should give organizations that deploy AI significant leeway to implement AI to perform security functions. Allowing users to opt-out of security systems can materially harm many users across the network. If one user is compromised because they opted-out of advanced security controls, that can create risk for other users or the system itself. We see this today when users turn off important security features like auto-updates, resulting in device compromise which in turn leads to follow-on exploitation. Additional opt-out requirements would only exacerbate the problem. While safeguards may be necessary to ensure the AI system's purpose is security, regulatory approaches must consider system-wide impacts from restricting AI security tooling. This includes those which provide scores to the relevant entity around whether a given interaction is likely to be fraudulent, malicious, or compromised.

Promote, rather than prohibit, AI-powered security for critical infrastructure and public sector networks

Public sector and critical infrastructure organizations are highly targeted by malicious actors and require constantly improving defenses — yet new regulatory proposals and existing procurement practices limit the ability of these organizations to rapidly deploy new commercial solutions. Some of the greatest attacker innovations were developed to target and compromise these high-risk systems. Given the threat environment, defenders of these systems must be able to adopt state-of-the-art security tooling and practices, which will increasingly be AI-based. In time, AI defenses will be necessary, and possibly required, for critical systems. However some policy proposals take the opposite approach, seeking to restrict the use of AI in high-risk systems. Governments should ensure that new AI policies do not inhibit market access for AI security innovations. Existing procurement regimes on which many AI tools are rooted, such as those for cloud services, should be updated to ensure these technologies do not face significant bottlenecks. Otherwise, barriers to adoption will deepen technological asymmetries in favor of attackers.

Advance research cooperation to generate scientific breakthroughs

To realize AI's potential, government, industry, civil society and academia must come together to unlock advances to supercharge defenders. Each week, new research is published detailing novel attacks on AI systems. While this research is vital, we need more basic and applied advancements in how to protect AI systems, and how to use AI to protect classic systems.

Pursue research in key areas

We see multiple areas in two broad categories where fundamental advancements are needed:

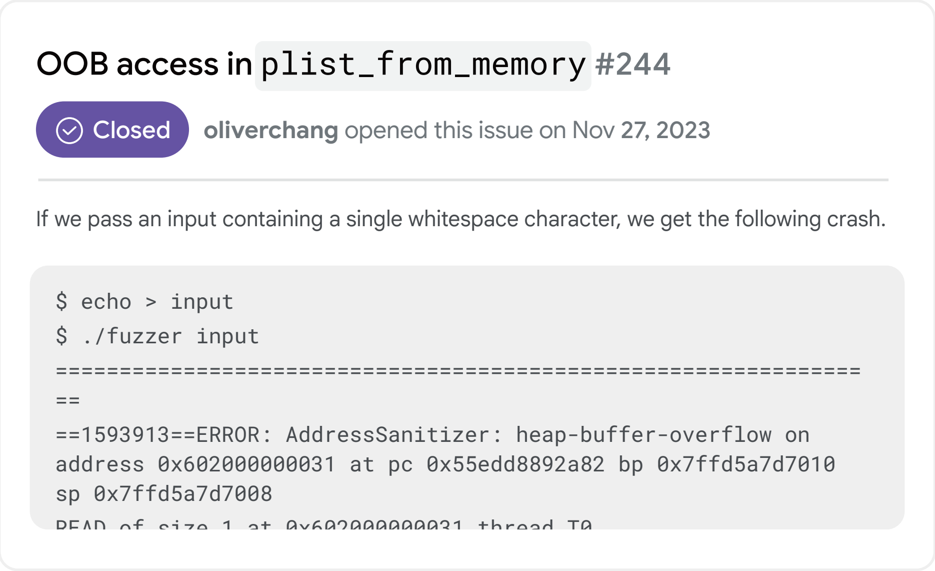
- **System safety in design and build**
  We've demonstrated some early successes in using AI to discover and fix security flaws faster by augmenting current techniques such as fuzzing. However, we need to go beyond doing so after the fact with research into how we can augment and accelerate all aspects of the security lifecycle.

  Scanning and fuzzing after code is written is good, but systems comprise more than just code, and catching systemic problems as they are designed and built is essential. AI-powered development tools can guide engineers towards code and configuration that is secure by design, and verify formal properties of a system to ensure that it does not create new safety, privacy, or compliance risks.

  Keeping design artifacts such as documentation, reviews and assessments up to date is a perennial problem, despite all of the promises that they will be living documents. AI can help here as well: being able to reason about both a system and its documentation, and find or fix inconsistencies, makes those artifacts more trustworthy, and frees up human attention and effort. This will be especially important at large scales, where the security of "systems of systems" requires looking at more than just the constituent parts.

- **System safety in use**
  Today, defenders can be overwhelmed by the "needle in the haystack" problem: collecting vast amounts of data is often easier than interpreting it. AI can make significant headway against this; AI models excel at sifting through huge masses of data to detect, understand, and ultimately respond to patterns of unusual or malicious activity, whether or not they were previously known, from threat intelligence and hunting to incident response.

  Research is needed on new techniques that can be used across the operational lifecycle to detect threats, synthesize and deploy mitigations, and then document and inform the people who are ultimately responsible for the safe operation of a system. Research into where human feedback is most effective will also be very important.



OOB access in `plist_from_memory` #244

oliverchang opened this issue on Nov 27, 2023

If we pass an input containing a single whitespace character, we get the following crash.

```
$ echo > input
$ ./fuzzer input
===============================================================
==
==1593913==ERROR: AddressSanitizer: heap-buffer-overflow on
address 0x602000000031 at pc 0x55edd8892a82 bp 0x7ffd5a7d7010
sp 0x7ffd5a7d7008
READ of size 1 at 0x602000000031 thread T0
```

Figure 4

This bug was found by using Gemini to write new fuzz tests for open source projects, leading to coverage increases of up to 30% across more than 120 projects.

In addition to these areas, we also see the need for even more forward-looking research:

- **AI Agents for Security**
  AI-based agents hold great promise for managing some kinds of security issues. How to build these agents, measure and monitor their performance and accuracy, and explain their actions is an active research area.

- **Novel threats and solutions**
  Every technological advance brings entirely new classes of both threats and solutions. Beyond all of the applications of AI in areas we know about, we need more exploratory and speculative research into new types of threats and new capabilities for defense. In particular, some characteristics of new AI systems may result from particular implementation techniques, but others may be more inherent. Designing controls, quality measurement, and reliability mechanisms require research in their own right, including how to apply AI itself to these questions (for example, Reinforcement Learning with AI Feedback (RLAIF)).

Conclusion

This is our once-in-a-generation moment to change the dynamics of cyberspace for the better — a chance for profound transformation, not incremental gains.

While we must build AI to be safe and secure, the technology already shows tremendous promise to address some of the greatest risks we face online.

Today, AI is helping us detect threats and reduce toil and burnout for defenders.

Tomorrow, if we work together, we believe AI can raise organizations up to a capable level, help us address the complexity crisis that is the source of countless breaches, and render attacker tactics obsolete.

As threats continue to multiply, exacerbated now by attacker use of AI, we have no choice but to seize this moment. Through partnership and focused investments, we can reverse the Defender's Dilemma.

Appendix

Appendix A

Roadmap for Reversing the Defender's Dilemma

Prepare and prevent

ATTACK SURFACE

Defenders

Current state

□ 🟨 □

Complex and often legacy infrastructure is hard to defend. Resources are necessarily static and discoverable. Defenders have limited knowledge of exploits, must react to adversarial techniques when disclosed, and manually develop mitigations (patching vulnerabilities, updating configurations).

Future state

🟩 □ □

AI helps manage infrastructure complexity with fewer vulnerabilities. More standardization and secure-by-design and default principles makes it easier to defend. AI systems learn from global attack data and find vulnerabilities more comprehensively than attackers. Fixes can be automatically applied.

Attackers

Current state

□ 🟨 □

The attack surface provides limited visibility for attackers. Attacker resources are obscured and often mutable. Attackers can find vulnerabilities and keep them private to exploit many organizations undetected.

Future state

□ □ 🟥

AI helps manage organization's attack surface, providing fewer opportunities for attackers. Attackers have difficulty developing new exploits because AI systems can find and fix them faster.

KEY ENABLERS

Standards: Require use of existing, multi–stakeholder standards and secure software development practices for AI systems to reduce system complexity, vulnerabilities, and standardize defense needs.

Deployment/Adoption: Rapid procurement and deployment of new innovations by defenders.

Research and Development: AI assistance for attack surface management and related controls; secure coding practices; vulnerability discovery and remediation.

Data: International framework that preserves the ability of AI systems to learn from global incident data and operate across borders.

SPEED OF INNOVATION

Defenders

Current state

□ □ 🟥

Organizations iterate more slowly than attackers, and face resource and regulatory constraints to innovation.

Future state

🟩 □ □

Organizations benefit from AI solutions which can apply learnings and update technology on the user's behalf.

Attackers

Current state

🟩 □ □

Few barriers constrain attacker use and development of new, innovative tools. Attackers can change strategies quickly with lag in defender response.

Future state

□ □ 🟥

While attackers can use new tools, their access to AI technology is worse than defenders, and they quickly face agile mitigation strategies.

KEY ENABLERS

Deployment/Adoption: Rapid procurement and deployment of new innovations by defenders.

Data: International framework that preserves the ability of AI systems to learn from global incident data and operate across borders.

Education: A cyber workforce that is trained to make best use of AI technologies.

Research and Development: Focus on both one-to-many defenses, one-to-one detection and defense.

Appendix A

Roadmap for Reversing the Defender’s Dilemma

Detect and respond

DETECTION

Defenders

Current state

⬜
⬜
🟥

A murky and dynamic threat landscape means defenders have incomplete information and do not know where attacks will come from next. Defenders face difficulties attributing attacks.

Future state

🟩
⬜
⬜

AI-integrated defensive systems with access to telemetry and analysis can aid human responders and generate automated actions. AI systems can scale no matter the level of alerts and complexity of environment. AI systems can aid with attribution.

Attackers

Current state

🟩
⬜
⬜

Attackers study attack surface for as long as they like. They can choose to attack at any moment and from any vantage, as frequently or infrequently as desired. Attackers can obfuscate their activities, making detection and attribution difficult.

Future state

⬜
🟨
⬜

Attack surfaces and infrastructure cannot be easily studied or are hardened. Attackers cannot effectively obfuscate the source of attacks.

KEY ENABLERS

Autonomous defense: Nuanced rules enabling automated incident response with effective human oversight.

Iteration: Protect the ability to learn and train AI quickly. Documentation/testing balanced against the need for rapid evolution.

Data: International framework that preserves the ability of AI systems to learn from global incident data and operate across borders.

Opt-out exceptions: Prevent opt-outs on security to ensure that system data is complete for analysis / defense.

Research and Development: Using historical breach and incident data to train models; finding commonalities in zero-day exploits. Research on making attacks computationally expensive with AI, e.g. autonomous hardening and obfuscation.

Partnership: International cooperation, information sharing, investigation of threats and best practices.

RULES OF ENGAGEMENT

Defenders

Current state

⬜
⬜
🟥

Legal and ethical constraints impede response.

Future state

🟩
⬜
⬜

Legal and ethical approach to employment of AI defenses is well-thought out and builds trust in its use. Defenders benefit from strong relationships with law enforcement and investigatory entities.

Attackers

Current state

🟩
⬜
⬜

Few legal impediments block malicious activities.

Future state

⬜
⬜
🟥

New restrictions make it harder for adversaries to access innovative AI technology.

KEY ENABLERS

Guardrails: Align AI security technologies, including autonomous agents, with our values and ensure effective human oversight.

Standards: Require use of existing, multi–stakeholder standards and secure software development practices for AI systems to reduce system complexity, vulnerabilities, and standardize defense needs.

Partnership: International cooperation, information sharing, investigation of threats and best practices.

Research and Development: AI to assess development artifacts; AI to authenticate good actors.

Sanctions/Embargoes: Prevent adversarial state actors from getting the most advanced technologies, including infrastructure.

Appendix A

Roadmap for Reversing the Defender’s Dilemma

Access to AI

COMPUTE

Defenders

Current state

🟨

Access to resources depends on size of organization.

Future state

🟩

Access to compute resources will still depend on size of organization, but AI systems are baked into widely-used platforms providing benefit to all organizations.

Attackers

Current state

🟨

Access to compute is commensurate with their size and reach, but may require significant resources.

Future state

🟥

Ability to acquire compute is limited, and broadly accessible resources have guardrails in place.

KEY ENABLERS

Restrict access: Control access to data, compute and other resources to train advanced AI.

Partnership: International cooperation, information sharing, investigation of threats and best practices.

DATA

Defenders

Current state

🟨

Defenders often have ample access to data resources, but may not have strong data sharing partnerships.

Future state

🟩

The defensive community benefits from ample access to data resources, including specialized data about attacks and infrastructure across other networks and internationally. Specialized training data sets exist for defender use.

Attackers

Current state

🟨

Access to data is commensurate with attacker’s size and reach, but lack organized data sharing initiatives

Future state

🟨

Attackers do not have access to data at the scale of defenders. The most useful information may be limited.

KEY ENABLERS

Data: International framework that preserves the ability of AI systems to learn from public information and global incident data, and operate across borders.

Partnership: International cooperation, trusted information sharing, investigation of threats and best practices.

Restrict access: Control access to data, compute and other resources to train advanced AI.

Research and Development: Determine which synthetic datasets are helpful; find how datasets can be effectively combined.

Appendix B

Google's Secure AI Framework

AI is advancing rapidly, and it's important that effective risk management strategies evolve along with it. To help achieve this evolution, we introduced the Secure AI Framework (SAIF), a conceptual framework for secure AI systems. **SAIF has six core elements:**

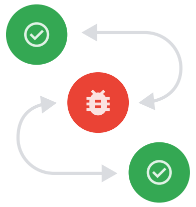
Expand strong security foundations to the AI ecosystem

Leverage secure-by-default infrastructure protections and expertise built over the last two decades to protect AI systems, applications and users. At the same time, develop organizational expertise to keep pace with advances in AI and start to scale and adapt infrastructure protections in the context of AI and evolving threat models. For example, injection techniques like SQL injection have existed for some time, and organizations can adapt mitigations, such as input sanitization and limiting, to help better defend against prompt injection style attacks.

Extend detection and response to bring AI into an organization's threat universe

Detect and respond to AI-related cyber incidents in time by extending threat intelligence and other capabilities. For organizations, this includes monitoring inputs and outputs of generative AI systems to detect anomalies, and using threat intelligence to anticipate attacks. This effort typically requires collaboration with trust and safety, threat intelligence,and counter abuse teams.

Automate defenses to keep pace with existing and new threats

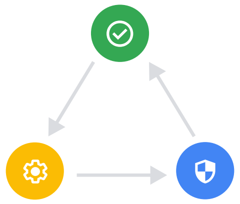


Harness the latest AI innovations to improve the scale and speed of response efforts to security incidents. Adversaries will likely use AI to scale their impact, so it is important to use AI and its current and emerging capabilities to stay nimble and cost effective in protecting against them.

Harmonize platform level controls to ensure consistent security across the organization

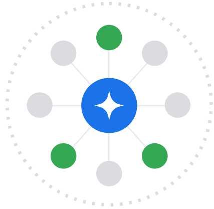
Align control frameworks to support AI risk mitigation and scale protections across different platforms and tools to ensure that the best protections are available to all AI applications in a scalable and cost efficient manner. At Google, this includes extending secure-by-default protections to AI platforms like Vertex AI and Security AI Workbench, and building controls and protections into the software development lifecycle Capabilities that address general use cases, like Perspective API, can help the entire organization benefit from state of art protections.

Adapt controls to adjust mitigations and create faster feedback loops for AI deployment



Constantly test implementations through continuous learning and evolve detection and protections to address the changing threat environment. This includes techniques like reinforcement learning based on incidents and user feedback, and involves steps such as updating training data sets, fine-tuning models to respond strategically to attacks, and allowing the software that is used to build models to embed further security in context (e.g. detecting anomalous behavior). Organizations can also conduct regular Red Team exercises to improve safety assurance for AI-powered products and capabilities.

Contextualize AI system risks in surrounding business processes



Conduct end-to-end risk assessments related to how organizations will deploy AI. This includes an assessment of the end-to-end business risk, such as data lineage, validation and operational behavior monitoring for certain types of applications. In addition, organizations should construct automated checks to validate AI performance.

This report includes extensive research from dozens of sources and comes in print and online versions. The online version contains links to relevant sources.