LEVELING THE **PLAYING FIELD**

As criminals expand their use of AI in cyberattacks, the information security field is fighting back in kind—but can AI really defeat AI?

AI is everywhere. It is in places it should be, smoothing the customer service experience, detecting fraud in financial institutions, boosting healthcare processes to save more lives. It is also in places it shouldn't be, popping up to mistakenly replace perfectly solid processes when executives get unduly excited, or polluting the web with worthless, often incorrect content. And AI is in places it *really* shouldn't be: in the hands of cybercriminals, eager to use it to sharpen their attacks, gather more data, and collect more ransoms.

Every facet of AI is at play, from generative AI to machine learning and deep learning, creating new malicious tools and techniques, invading and poisoning networks, and disrupting legitimate AI tools. The attacks cybercriminals have relied upon for many years have been rejuvenated by AI, made stronger and more effective. Information security is undergoing a slow reset, where new methods of executing old attacks are replacing the reliable techniques that once safeguarded the world's networks.

Phishing 2.0

Phishing might be the most well-known attack vector, and it is definitely the most plentiful. It is said to account for a considerable portion of all emails, with some sources claiming (but not proving) that billions† of phishing emails are sent daily. Whatever the true number, phishing works. IBM's latest X-Force Threat Intelligence Report states that phishing is responsible for 41% of security incidents. CSO magazine, writing at phishing's peak at the height of the pandemic, reported a much higher number, over 80%.

60% OF RECIPIENTS FALL VICTIM TO AI-GENERATED PHISHING EMAILS†

The human element is the most vulnerable part of digital defense, after all. One tired employee, one rushed executive, one distracted individual with their guard down is all it takes to capture vital credentials, giving criminals fuel for the next stage of their attack. In the age of AI, however, the numbers don't paint the whole picture.

Phishing is predominantly a practice of throwing everything against the wall and seeing what sticks, but the world has become wise to scattergun phishing. Our email defenses are built strong, and little of the daily wave of phishing attempts makes it through. The Nigerian princes of the 1990s taught us well, so even the least tech-savvy employee will likely smell a rat if a random phishing attempt hits their inbox—but the same cannot be said of carefully crafted spear phishing emails. Designed to look and feel authentic, probably claiming to come from someone in a position of power, and possibly even originating from a previously compromised account, these are a different breed of attack altogether.

AI's potential influence here is likely rather obvious. Criminals use generative AI to skip most, if not all, of the hard work of preparing spear phishing attacks. They can train their models on publicly available data like social media posts or organizational charts, or even on breached data, known or unknown. AI can then generate highly realistic emails cognizant of everything from interpersonal relationships to the formatting, style, and writing tics of an individual.

The rise of deepfakes

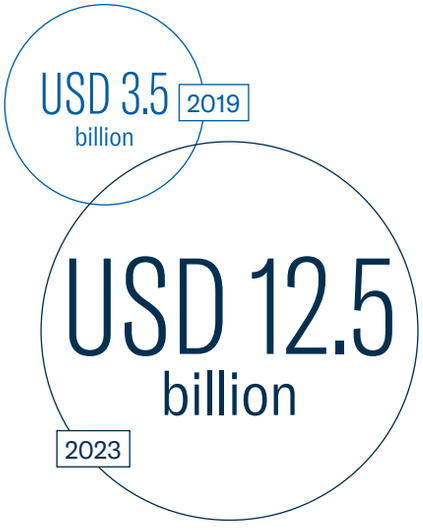
Text is not all that can be emulated, as the headlines have made all too clear. Generative deep learning AI trained on voice and video enables attackers to take a new route, striking via synthetic media—specifically, by so-called deepfakes. They might deploy phone calls supposedly from a friend, colleague, or family member, using information from social media to make conversations more personal and convincing. They can issue eerily realistic videos, mix the two, or even build filters to generate deepfake content live and on the fly.

The Department of Homeland Security sums up the threat of deepfakes, stating that it "comes not from the technology used to create it, but from people's natural inclination to believe what they see." Today's attempts are noticeable to the trained eye, but the DHS suggests that "the severity and urgency of the current threat from synthetic media depends on the exposure, perspective, and position of who you ask." The right fake in front of the wrong person could easily be the first step of a massive breach.

The threat is already here. Deep learning AI techniques are leagues ahead of their abilities and only getting more sophisticated. As the availability of high-end processing rises and its cost

lowers, the threat level, frequency, and convincing nature of deepfakes will increase—there is nothing as effective as confidence in encouraging a victim to drop their guard.

AGGREGATE **CYBERCRIME LOSSES** ROSE FROM USD 3.5 BILLION IN 2019 TO USD 12.5 BILLION IN 2023†



Adaptive tactics

Equally as concerning as the human element, AI has begun dissolving the defenses that the information security sphere has built over the past few decades. Criminals no longer need to spend their time probing a network, because an AI can be trained to automatically analyze an institution's digital footprint. Automated reconnaissance allows potential attackers to discover vulnerabilities, map network structures, and essentially be presented with a plan of attack with little to no effort on their part.

Applied to the creation of malware, AI is equally insidious. Consumer-grade transformers like ChatGPT can, with the correct prompts, be coerced into writing malicious scripts. The controls which bind the output of chatbots are entirely circumventable; if a hacker wishes to create a damaging and undetectable PowerShell script, ready for deployment, they can.

Along similar lines, AI can design polymorphic malware, which changes its code or behavior to evade traditional antivirus detection methods, hugely increasing the level of remediation required to keep it at bay. It is even possible to use AI to coordinate DDoS attacks, optimizing and scaling them once it identifies the most effective vectors and patterns to overwhelm a system.

And cybersecurity professionals can't forget that AI is a two-way street. As workers rely more and more on AI to perform their day-to-day tasks, they also input increasing amounts of sensitive data into AI engines. That's an incredible amount of trust placed in a tool that may be spoofed, intercepted or insecure, a tool we already know can be tricked and exploited. Few companies have the resources to run a fully internal GPT, much less one with the power of the leading engines, meaning any data input into an AI engine is essentially being willingly placed out in the open.

Poisoning the well

For companies using AI to their advantage, it is important to realize that the AI battle is not one fought along a defined front. A criminal acting in enemy territory can disrupt the AI systems organizations are coming to rely on. By poisoning a data set, for instance, corrupted data could cause flawed or biased decision-making or encourage a public-facing AI to output undesirable responses. If an exploited AI is responsible for automation, it could cause a cascading failure before it is even noticed.

We've seen high-profile examples of poisoned AI before. One of Microsoft's earliest chatbots, Tay, was designed to develop conversational understanding by talking with humans on Twitter, learning from their posts. Even in 2016, Twitter offered a rather corrupt data set. Tay lasted just one day on the platform, which was all it took for a coordinated attack by a subset of Twitter users to teach the AI to be racist, antisemitic, and lewd.

Amidst the AI gold rush, many businesses are more interested in pulling the trigger on new AI innovations than checking that the safety is on. Microsoft's apology statement of 2016 reverberates today, almost a decade later: "To do AI right, one needs to iterate with many people and often in public forums. We must enter each one with great caution and ultimately learn and improve, step by step."

THE **GLOBAL COST** OF CYBERCRIME IS ESTIMATED TO REACH ALMOST **USD 14 TRILLION** BY 2028†

If AI is deployed without the proper safeguards in place, hackers need not even poison the data set. So-called adversarial attacks feed such systems maliciously crafted data, designed to manipulate them into making errors or breaking entirely. Even public-facing transformers have their quirks. It was recently discovered that ChatGPT's generation fails if it is asked to output certain people's names, for example—fine if one's use of AI is frivolous, but a disaster in a production context.

AI advantages

Logically, the information security sector must fight back with its own AI tools. There is nothing moving at quite the same pace nor growing in complexity in quite the same way as the world of machine learning and generative AI. However, dealing with the AI onslaught is not a task that IT professionals can complete manually.

Predictive models

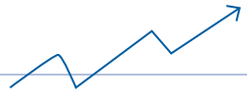
The ability of AI systems to analyze vast amounts of data in real-time means they are well suited to identifying patterns and detecting anomalies that indicate possible cyberthreats. Predictive models may offer a glimpse into the future, helping to prevent future attacks based on prior knowledge. In essence, AI takes on the role of the seasoned information security professional—the person who always knows the answer. The more data AI can collect, the more seasoned it becomes.

Behavioral analytics

By flagging deviations from usual patterns, behavioral analytics can identify unauthorized access and even work to fight against insider threats. If a user with legitimate credentials behaves strangely—accessing things they typically wouldn't or logging on at unusual times—AI can detect this and trigger an automation to stop it.

Specialized AI

Used to spot inconsistencies in AI-generated media, specialized AI helps uncover deepfakes and reinforce trust in communications. AI can detect polymorphic malware through heuristic analysis and pattern recognition. It can sanitize data and inputs to preserve the integrity of AI models, analyze traffic to spot unusual activities, and automatically trigger an incident response if an intrusion is detected. With AI, information security efforts can fire as incidents happen rather than reacting after the fact. However, AI is not a complete answer—certainly not a drop-in solution.



THE AI CYBERSECURITY MARKET WILL GROW TO OVER **USD 60 BILLION** BY 2028†

An expanded defensive posture

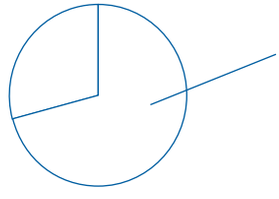
However difficult the process of developing and implementing them may be, AI-driven cybersecurity solutions must be part of the defensive toolkit. The scale of AI activity is simply too great. Those that ignore the unique advantages of AI tools will inevitably struggle against the sophistication and frequency of AI-driven attack techniques. AI may even prove cheaper than more manual security processes, with IBM suggesting that companies using AI enjoyed an average saving of USD 2.22 million† against those that did not. That does not mean, though, that defensive AI is the one true solution to the onslaught of AI hacks.

The fact is criminals have an innate advantage. Their singular focus allows them to remain one step ahead of any immature AI information security strategy. Institutions must pour time and effort into monitoring and updating all AI systems, including the AI engines hastily integrated into day-to-day business practices, to detect and mitigate adversarial attacks. Even then, it is safe to assume that criminals have as-of-yet unknown attacks ready to deploy.

Defensive AI is one tool of many—effective in the right places, but not infallible, and arguably less important than behaving in a proactive manner and getting the fundamentals right. AI may be able to offer alerts and guidance, but it does not make humans less vulnerable as an attack vector. Conducting regular security awareness training helps hammer home the importance of behaving in a safe and cynical manner.

GENERATIVE AI COULD PUSH AVERAGE FRAUD **LOSSES** TO **USD 40 BILLION** BY 2027†

As it always has, strong information security relies on multi-layered defense strategies. AI is one of these, but institutions may not be able to weather the storm without a culture of strong policy making, intelligent access controls, and a Zero Trust model that helps ensure such policies work as they are supposed to. There is no sense in being dazzled by the new and exciting, because the fundamental principles of security have not changed at all.



71% OF ORGANIZATIONS HAVE UNFILLED CYBERSECURITY POSITIONS†